Diagnosing Semantic Properties in Distributional Representations of Word Meaning

Sommerauer, Pia Johanna Maria

2022

document version
Publisher's PDF, also known as Version of record

Diagnosing Semantic Properties in Distributional Representations of Word Meaning

Pia Sommerauer

The research reported in this thesis was funded by the Netherlands Organization of Scientific Research (NWO) via the PhD in the Humanities fund.
Cover photo by:
https://www.istockphoto.com/nl/foto/winter-sky-on-the-flour-gm598819912-102757177
Cover design by: Hanna Gerhardt
Printed by:
Ipskamp printing
https://www.ipskampprinting.nl/

VRIJE UNIVERSITEIT

DIAGNOSING SEMANTIC PROPERTIES IN DISTRIBUTIONAL REPRESENTATIONS OF WORD MEANING

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geesteswetenschappen
op woensdag 8 juni 2022 om 15.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Pia Johanna Maria Sommerauer

geboren te Graz, Oostenrijk

promotoren:	prof.dr. P.T.J.M. Vossen prof.dr. A.S. Fokkens
promotiecommissie:	prof.dr. A.J. Cienki prof. C.D. Fellbaum prof. A. Lenci prof. M. Nissim dr. A. Alishahi

Acknowledgments

Writing a PhD thesis is often thought of as a lonely process. In the past four and a half years, however, I have been lucky enough to discover that it doesn't have to be. I have had the privilege to encounter, learn from, and work with highly motivated, inspirational, but above all, exceptionally kind people. While I want to consider myself driven by curiosity and a passion for research, the truth is that I never would have finished a single experiment (let alone written a single page) without their support, inspiration, mentorship, and friendship.

My PhD journey started before the first day of my PhD. Piek Vossen helped me to prepare a PhD application for an individual PhD project. From day one, Piek made it clear that I was in charge of my project and could decide what to investigate. Throughout my PhD, Piek offered support and advice on a scientific level, but, perhaps more importantly, kept me inspired with his visionary perspective and excitement for language. Above all, though, I've come to know Piek as a supervisor and boss who is immensely loyal to his team. The playful attitude and lightness he brings to research and his overall enthusiasm are truly infectious!

Officially, my 'academic' journey started when Antkse Fokkens hired me as a student assistant to work on an interdisciplinary research project about detecting stereotyping in natural language. From the start, I've come to know Antske as an exceptional supervisor. Antske has the ability to listen to sometimes naive ideas and see a potential 'spark' in them. In the past years, I've had the privilege to get to know Antske as a colleague and friend and have seen her bring this open, non-judgmental attitude to all of her collaborations. I am grateful to have had an academic mentor who so clearly stands for teamwork and fairness.

Next to my official supervisors, I have had a number of unofficial supervisors, teachers, and mentors among my colleagues at CLTL. Thanks to Marten Postma and Emiel van Miltenburg, who famously once put together a Python course overnight, I've been able to learn programming up to a level I can teach it myself (so they don't have to anymore). I've also learned a great deal (from basic Latex skills to dealing with existential crises) from my PhD 'siblings' Chantal van Son, Minh Leh, and Filip Ilievsky. Thank you to Hennie van der Vliet, Roser Morante, Isa Maks, Tommaso Caselli, Sophie Arnoult, Paul Huygens, Marieke van Erp and all the others who have created a warm and welcoming atmosphere at CLTL.

Since more recently, I've also been able to learn from my newer colleagues. Their critical questions and diverse interests are a true source of inspiration! Levi Remijnse, Selene Baez Santamaria, Baran Barabarestani, Lisa Beinborn, Angel Daza, Jonathan Kamp, Lea Krause, Taewoon Kim, Urja Khurana, Jaap Kruijt, Myrthe Reuver, thank you for keeping me company during the last bit of my PhD and supplying me with chocolate and other much-needed 'thesis fuel'.

Next to immediate colleagues, I've had the opportunity to collaborate with colleagues outside of the lab. Arianna Betti, Yvette Oortjin, Jelke Bloem, and Francois Meyer took me on an adventure through philosophical writings and out of the comfort zone of my own

discipline. Bettina Speckman, Pantea Haghighatkhah, and Kevin Verbeek allowed me to explore semantic spaces through the lens of geometry. Thank you for your trust and your online company during a period of intermittent lockdowns. Thank you to Sanne Hoeken for getting involved in my research far beyond the level expected from master's students.

Submitting a PhD thesis is one tiny moment at the end of a long process. Much of this process started long before the first day of my PhD. I would like to thank my mentors who taught me that seeking out challenging tasks can be rewarding and encouraged me to pursue my academic interests. Thank you to Pia Jaritz, Małgorzata Fabiszak, and Nikolaus Ritt.

Pursuing a PhD in the Netherlands meant that I've spent the past years away from home. I'm grateful for my friends in the Netherlands who helped me create a new home in Amsterdam. At the same time, it was incredibly comforting that my friends in Austria remained so present in my life. Thank you for your long-lasting long-distance friendship.

Computational linguistics is perhaps not a very obvious career choice (or at least it was not when I started out). Even less so were my undergraduate studies in which I could not decide whether I wanted to earn a degree in English literature, linguistics, philosophy, or law. I would like to thank my parents for trusting my intuitions and encouraging me to pursue my interests, regardless of whether they seemed to lead to a 'safe' career.

Finally, thank you to Georg for the endless support and patience. Much of what got me through the last years can probably be traced back to your openness and optimistic attitude, your willingness to listen to endless monologues, and your ability to turn the most annoying problems into entertaining puzzles.

Amsterdam, Spring 2022

Contents

Introduction

Computers, Language, and Word Meaning

With the availability of high computing power and large amounts of data, automatic systems that process language have become increasingly powerful. It is possible to extract different aspects of information from texts and use automatic systems to perform complicated linguistic tasks. For many, the most impressive achievement might be that automatic systems have started to reach high performance on tests that assess a system's ability to engage in common sense reasoning. Consider the following example:

(1) Sam pulled up a chair to the piano, but it was broken, so he had to stand instead. What was broken?[1]

In order to answer the question posed in Example 1 correctly, it is necessary to infer which of the candidate referents in the sentence are likely to be referred to by the pronoun *it*. Possible referents are Sam, the chair, and the piano. If the name *Sam* is recognized as a likely reference to a human, two candidates remain: *chair* and *piano*. In order to decide which of the remaining candidates is the more likely referent, it is crucial to know that an essential function of a chair is to provide a surface for sitting and that having to stand means not being able to sit. If Sam has to stand, he cannot sit on the chair and thus it is likely that the thing that is broken is the chair rather than the piano. A computational system that can solve this problem can thus be expected to have a rich, semantic representation of the words *chair* and *stand* and use them to reason over the possible referents of the pronoun *it*.

The interpretation of words is a crucial component of many Natural Language Processing (NLP) problems (traditionally referred to as 'tasks'). For instance, the correct resolution of pronouns in a text may require rich, semantic knowledge. Other highly semantic tasks also require deep semantic understanding, such as for instance question-answering or tasks that involve reasoning over specific entities or events expressed in texts (e.g. entity and event co-reference). Most state-of-the-art NLP systems rely on word representations that are derived from large amounts of textual data. While they lead to successful results for many tasks, they are by no means perfect and still make seemingly silly mistakes (e.g. Staliūnaitė and Iacobacci, 2020). This may indicate that the word representations they rely on are not always accurate. Beyond the problem of accuracy, it has been shown that word representations capture social biases that impact the behavior of systems and can lead to results that reproduce stereotypes (Zhao et al., 2019, 2018; Rudinger et al., 2018). Rudinger et al. (2018) show that co-reference resolution systems make mistakes rooted in gender bias. For example, female pronouns and words such as *surgeon* are likely not recognized as labels of the same referent

[1] Taken from the Winograd Schema Challenge `https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html`

even if the relation is expressed in the text (e.g. *The surgeon couldn't operate on her patient: it was her son.*).

In this thesis, I will explore the semantic content of data-derived representations that underlie many NLP systems. Beyond having practical implications for the field of NLP, the question of what aspects of word meaning can be captured purely on the basis of how they are used in texts addresses fundamental questions within the field of distributional semantics (Lenci, 2008). While this thesis does not provide exhaustive answers, it constitutes an approach towards 'diagnosing' different aspects of semantic knowledge in word representations used in high-performing NLP systems.

You shall know a word...

What does it mean to know the meaning of a word? Is it being able to recite its dictionary definition? Or does it mean being able to point to instances of the concept in the world? Being able to translate it into different languages? When do we know that a computational system 'knows' the meaning of a word?

While this question of how the meaning of a concept can be defined is notoriously difficult to answer, it is possible to think of various 'tests' that can indicate whether a person (or a computational system) has some sort of understanding of a concept. For instance, it is possible to assess knowledge about words in terms of similarities and differences using the following task:

(2) Which word pair is more similar?
   a. *strawberry - raspberry*
   b. *strawberry - bicycle*

There are multiple ways by which one could arrive at the correct answer: Both words in pair (a) refer to fruits whereas *bicycle* in pair (b) refers to a vehicle. Both words in (a) refer to red berries, which only applies to one of the words in (b). Both words in (a) refer to edible things, which is not the case in (b). All of these approaches lead to the correct answer, but use different aspects of word meaning to reason over the similarities and differences between the words. The correct answer by itself does not reveal which approach was taken.

A more explicit approach is to ask for an enumeration of all aspects of the meaning of a word in the form of features. For instance, the word *strawberry* could be defined by **red**, **fruit**, **berry**, **edible**, **juicy**, **sweet**. The difficulty of such an approach lies in assessing the answer: When do we have enough features to determine that the answer is 'correct'? How fine-grained do the features have to be? Do they only have to include salient or discriminatory aspects? Where is the line between salient features and non-salient features?

Despite their problems, semantic features allow us to determine whether a model can capture specific aspects of word meaning. Rather than attempting to produce exhaustive descriptions in terms of feature lists, we can use them in a more targeted way: Given a collection of words, is it possible to test whether a computational model can detect all words with a specific feature?

(3) Which of the following concepts can be described by **can fly**?

a. *seagull*

b. *table*

c. *airplane*

d. *penguin*

e. *bee*

f. *strawberry*

Identifying all words whose meaning can be described by **can fly** entails knowing that things referred to as *seagull*, *airplane* or *bee*, despite belonging to different categories, all share the ability to fly. Beyond this, it also entails knowing that neither of the things described by *table*, *penguin* and *strawberry* have the ability to fly. Thus, this task cannot be solved by relying on general similarity: Even though *seagull* and *penguin* are similar to each other in many ways, they do not share the ability to fly. *Airplane* and *bee*, in contrast, describe radically different things that share hardly any features except for the target feature.

...by the company it keeps?

A different approach to word meaning argues that knowing the meaning of a word in essence means knowing how to use the word correctly. From this perspective, word meaning arises from the different situations in which a word occurs. Intuitively, we can infer the meaning of a new word based on its surrounding words. In the following examples, a word has been replaced by the character 'X', but can still be inferred fairly easily based on the words in its surroundings:[2]

(4) a. Would you like a drink, or X?

b. The X wasn't very hot though, made in a filter pot, but it was good.

c. Eugene put a spoonful of powdered X into his cup and then filled it with hot water.

d. She sees that there is a cup of steaming hot X awaiting him and the two chat informally as she presents the rules of the center and explains procedures.

e. She could not face X or tea without milk, and was always craving types of food that were not available aboard a sailing ship.

Most people will be able to infer that X must refer to a hot drink that is similar to tea, often consumed with milk, and can be prepared by means of a filter or dissolving powder. For competent speakers of English familiar with the target concept, it is most likely easy to guess that the word replaced by X must most likely be *coffee*, as they are able to interpret the words in its environment and probably remember that they have seen *coffee* appear in similar linguistic environments.

[2]The example sentences are taken from the Brown corpus (Kučera and Francis, 1967).

The intuition of being able to infer word meaning based on word co-occurrence has been formalized in the Distributional Hypothesis (Harris, 1954; Firth, 1957) and implemented as a computational approach to the representation of word meaning. In such computational models of meaning, individual words are represented in terms of how often they co-occur with other words in the vocabulary. Such representations can either be created by means of vectors that contain co-occurrence counts over the entire vocabulary or by means of machine learning models. Regardless of the specific computational architecture underlying the model, the intuition remains the same: Each word is represented by a vector in a vector space. Words which appear in similar contexts receive similar vector representations and thus appear in close proximity to one another in the vector space. The words *coffee* and *tea* should be placed close together, just like the words *strawberry* and *raspberry*. In contrast, the words *raspberry* and *bicycle* should be placed in different areas of the semantic space.

Such word representations have not only been shown to reflect human similarity judgments to some degree, but, more importantly, have proven highly successful when used in systems designed to perform complex semantic tasks. For instance, such distributional word representations have led to high performance for many NLP tasks, such as sentiment classification (i.e. automatically detecting whether a text expresses a positive, negative or neutral attitude towards something) (Socher et al., 2013), co-reference resolution (Zhou and Xu, 2015), and named entity recognition (Pennington et al., 2014). The fact that embedding representations improve performance on such tasks implies that they somehow capture rich, semantic information that can be used by systems to make semantically informed decisions. Experiments that compare distributional spaces to spaces defined by human-elicited semantic features show that they do, at least to some extent, represent comparable information (Fagarasan et al., 2015, e.g.). However, it is not clear whether they can capture specific semantic features that can be used to reason over fine-grained differences between concepts.

Such semantic representations can be created by machine learning models that learn to predict word-context combinations. Such models can be seen as language models that provide word representations by means of aggregating information over all contexts of a word in a corpus in a single representation. A different way of modeling language is to 'learn' contextualized representations by means of predicting the next word in a sequence. The intuition is easy to understand for anyone who has ever listened to someone on the phone with bad connection. Consider the following utterances:

(5) a. Hi, just wanted to know what you've been up [noise]
    b. I just got off work and am waiting for the [noise]
    c. It's been raining all [noise]
    d. What should we have for [noise]

Most competent speakers of English will be able to make fairly accurate guesses about which words have been cut-off by the bad connection. A traditional language model is trained on predicting the next word in a sequence of words. By performing this task, the model acquires certain aspects of linguistic information that arise from the distribution of words over different contexts. In contrast to the context-free distributional models described above, such models acquire context-dependent representations of words. Rather than creating a

single representation for each word, the models represent words given particular contexts. For example, such models should be able to capture that the word *star* can appear in contexts in which it refers to a celestial body and in contexts in which it refers to a celebrity.

Building upon this core idea, a family of recently proposed models (Bert and Roberta (Devlin et al., 2019; Liu et al., 2019)) have led to considerable performance gains for a number of tasks requiring semantic knowledge. Context-free and contextualized models differ in terms of their architectures and, more importantly, in terms of how they are used to perform NLP tasks. Rather than extracting word representations and using them in a downstream tasks (context-free models), the new variants allow for a set-up in which the entire trained language model is 'fine-tuned' to perform a particular NLP task.

Despite impressive performance gains, current systems are still far from perfect and make embarrassing mistakes. For instance, a systematic study of various linguistic phenomena shows that Bert can hardly ever process negation correctly (Ettinger, 2020). Using a question-answering set-up, Staliūnaitė and Iacobacci (2020) show that Bert and similar models tend to fail on examples that require inference, in particular when it involves compositional meaning (i.e. meaning that arises from the manner in which multiple words are combined to form a phrase or clause).

The successes and failures of language models raise questions about what aspects of linguistic, and, in particular, semantic knowledge they can capture. Are they even equipped for the type of reasoning they seem to be performing when, for example, solving complex pronoun resolution in so-called Winograd problems (Example 6)?

(6) The *trophy* doesn't fit into the brown *suitcase* because it is too **large**. What is too large? (Possible answers: the trophy, the suitcase)

Recent work on dataset biases has shown that many datasets aimed at testing the capabilities of language models, such as complex pronoun resolution (Sakaguchi et al., 2020; Elazar et al., 2021; Abdou et al., 2020) or Natural Language inferencing (Poliak et al., 2018) allow models to perform highly without actually solving the target task. Rather, the train- and test sets contain spurious correlations that models can exploit. These observations call into question to what extent language models can represent aspects of semantic knowledge.

Research Question

The goal of this thesis is to shed light on what aspects of word meaning language model representations carry. Knowing what information is there in the first place can help us understand what kind of information systems have access to and thus help us make better decisions about their design. Knowing what aspects of information tend to be encoded in distributional co-occurrence patterns is an important question in its own right as it may help us to arrive at a more fine-grained understanding of the distributional hypothesis. I formulate the central question of this thesis as follows:

What aspects of conceptual knowledge are reflected by the co-occurrence patterns captured by large-scale language models?

I address the central research question in three steps:

Step 1 : *Create **a model of conceptual knowledge and property expression** for the investigation of language model representations.*

Step 2 : *Capture human conceptual knowledge in a **dataset** suitable for the investigation of language models.*

Step 3 : *Use **interpretability methods** for context-free and contextualized language models to study which aspects of semantic knowledge they represent.*

Each of the three steps can be divided into sub-problems.

Step 1: A Model of Conceptual Knowledge and Property Expression

As illustrated above, determining whether a computational model has knowledge about specific semantic information is not trivial and comes with methodological challenges. An investigation of conceptual knowledge in distributional representations should lead to sound insights that can reveal, or at least indicate, general tendencies about the type of semantic information that tends to be reflected by co-occurrence patterns. I design a model of conceptual knowledge and property expression that fulfils two criteria:

- It should be possible to derive hypotheses about the specific factors that determine whether properties are expressed in corpora.
- It should be possible to represent conceptual knowledge in such a way that insights gained from the analysis of distributional models are informative given the **methodological challenges** of analyzing distributional models.

Step 2: A Dataset of Conceptual Knowledge

The investigation of conceptual knowledge in distributional representations requires a dataset that can be used to detect or 'diagnose' aspects of conceptual knowledge. Such a dataset should consist of properties and concepts that can be used for testing what aspects of word meaning large scale language models can reflect. As illustrated above, testing whether a computational model has knowledge about a specific semantic property is not trivial and comes with methodological challenges. In particular, it requires a specific distribution of positive and negative examples (recall the positive and negative examples of **can fly** discussed in Example 3, which included *seagull*, *airplane*, and *penguin*). Beyond testing knowledge about specific semantic properties, the dataset should be suitable for testing hypotheses derived from the *model of conceptual knowledge and property expression*. The construction of a reliable dataset that can lead to insights about property representation in language model representations entails the following steps:

Step 2-a Elicit a substantial number of fine-grained semantic judgments about conceptual knowledge from human participants.

Step 2-b Evaluate the quality of semantic judgments given that fine-grained semantic annotation tasks are likely to encompass phenomena that trigger disagreement between annotators.

Step 2-c Assess the resulting diagnostic dataset in terms of its ability to yield insights about specific hypotheses and its adherence to methodological requirements.

Step 3: Interpretability Methods and Experiments

The diagnostic dataset forms the basis for interpretability experiments designed for the analysis of non-transparent representations derived from machine learning models. Existing methods aim to provide insights into the black box nature of many current NLP approaches. While such methods have yielded insights about aspects of linguistic information captured by deep learning models, they struggle with a number of methodological challenges and are highly sensitive to noise (Belinkov and Glass, 2019; Belinkov, 2021). The model of conceptual knowledge and diagnostic dataset proposed in this thesis are primarily designed for the interpretation of context-free distributional representations. The emergence of contextualized models poses new challenges deriving reliable insights by means of diagnostic methods. I propose the following steps to approach these challenges:

Step 3-a Define control tasks and informative baselines that help to distinguish results caused by noise and accidental correlations from meaningful signals.

Step 3-b Use corpus analysis to verify the results of diagnostic experiments.

Step 3-c Design tasks that are suitable for the analysis of contextualized models.

Outline and Contributions

This thesis consists of five parts. Apart from Introduction and Conclusion, each part is divided into several chapters. In this section, I provide an overview of the chapters and their most important contributions.

Part I: Background

Part I presents the core concepts used in the thesis and outlines the most important findings from existing research about conceptual knowledge in distributional representations. I illustrate the limitations of distributional models through two use-cases that apply distributional representations to the study of specific concepts.

Chapter 1: Core Concepts and Related Work The chapter outlines the most important assumptions behind the core concepts underlying the research presented in this thesis: Firstly, it introduces the Distributional Hypothesis and different types of distributional semantic models. Secondly, it suggests semantic properties as an empirical approximation of word meaning and reviews existing datasets that capture this type of conceptual knowledge. Thirdly, the chapter reviews insights from existing research on conceptual knowledge in word representations.

Chapter 2: Two Use-Cases To demonstrate the limitations of embedding representations when used for fine-grained semantic analysis, I present two studies that evaluate the suitability of distributional semantic representations for the study of specific concepts from a usage-based perspective, in particular with respect to semantic variation and change.

Contributions

- The first use-case examines the methodological challenges of studying conceptual change using distributional models. The study resulted in a number of practical recommendations for drawing reliable conclusions from models prone to representing noise.

- The second use-case presents an evaluation of distributional models specifically designed for small data using a network of philosophical concepts. The results indicate that small-date models have potential, but cannot represent philosophical concepts accurately enough for supporting philosophical research.

Part II: Model

In Part II of the thesis, I introduce the model and design underlying the diagnostic dataset. The part addresses **Step 1** and consists of the following chapters:

Chapter 3: Semantic Property Information in Text Chapter 3 proposes a model of conceptual knowledge and property expression. A core assumption of the model is that semantic properties constitute highly implied knowledge. An explicit mention of a semantic property will, in many situations, constitute a violation of the maxim of quantity following Grice's (1975) co-operative principle. There are, however, a number of situations in which mentioning property-information can be justified. In some cases, properties are variable (e.g. apples come in different colors) and specifying a particular color can constitute necessary information. In other cases, properties enable particular activities or functions and determine how we interact with the world. Such activities and functions are components of events and thus likely to be mentioned explicitly (e.g. the sharp edge of a knife enables cutting). The model results in a framework of hypotheses about the expression of property information in text on the basis of specific relations between properties and concepts.

Chapter 4: Methodological Framework and Dataset Architecture The analysis of distributional representations is not trivial and runs risk of yielding misleading results. In Chapter 4, I consider the specific challenges involved in determining whether distributional representations carry information about specific semantic properties. Based on these considerations, I design a diagnostic dataset. Specifically, I select properties and concepts in such a way that they should lead to insights despite the methodological challenges. As illustrated above, distinguishing positive from negative examples of a property should only be possible if the property in question is identified (see positive and negative examples of **can fly** in Example 3).

Contributions

- Chapter 3 results in a theoretical model of the expression of semantic property-evidence in text. The model can be used to derive specific hypotheses about property expression.
- Chapter 4 presents a methodologically informed dataset design. Testing semantic property knowledge in distributional models runs risk of yielding misleading results. The design of the dataset presented in this chapter specifically addresses this challenge.

Part III: Dataset

Part III of the thesis is dedicated to the creation of a diagnostic dataset. It focuses on the collection of fine-grained and reliable semantic judgments by means of crowd annotation and the assessment of the resulting dataset. The three chapters of Part III outline the task design, annotation evaluation, and assessment of the resulting dataset. Each chapter addresses one component of **Step 2**:

Chapter 5: Annotation Task Chapter 5 outlines the collection of fine-grained semantic judgements of property-concept pairs by means of a crowd annotation task. The chapter presents the task design, annotation procedure, and approach chosen to recruit participants. The annotation was carried out in multiple cycles. I provide information about the versions of the dataset resulting from each annotation cycle.

Chapter 6: Evaluating Crowd Annotations Chapter 6 presents an approach towards evaluating crowd annotations. The diagnostic dataset presented in this thesis was collected by means of distributing the annotation efforts over many untrained workers who provided semantic judgments. To ensure that the task results in reliable annotations, it has to be evaluated. A common practice for such evaluations is measuring the agreement between participants. However, in the case of the diagnostic dataset, agreement alone is not a suitable indicator of quality. The task required participants to judge relations between properties and concepts. Participants were confronted with a number of linguistic phenomena that can trigger multiple justified interpretations. Thus, disagreement among annotators can be a reflection of linguistic phenomena rather than a sign of low quality. I design an alternative quality metric based on logical contradictions and evaluate it against expert annotations. The logic-based metric gives reliable insights about annotation quality. Even though the logic-based check I used was specific to the annotation task at hand, the principle of using a task-inherent, logic based metric rather than annotator agreement to establish quality can also be applied to other annotation tasks.

Chapter 7: A Corpus of Properties, Concepts, and Relations Chapter 7 presents an analysis of the diagnostic dataset. Based on the theoretical and methodological considerations, I constructed a dataset encompassing 21 semantic properties associated with positive and negative examples that can be used in diagnostic experiments. The set of examples is large enough to enable classification experiments using held-out test sets. Each property-concept

pair in the dataset was labeled with linguistic factors proposed in the model of conceptual knowledge by crowd annotators. Each semantic property dataset was evaluated with respect to the methodological requirements that informed the design.

Contributions

- The research presented in Chapter 6 resulted in an approach to the evaluation of semantic judgments made by crowd annotators. The approach is independent of inter-annotator agreement as it relies on the logic and coherence answers given by individual annotators. The evaluation metric provides information that is complementary to traditional inter-annotator-agreement metrics and the CrowdTruth approach presented by Dumitrache et al. (2019).
- Chapter 7 presents a diagnostic dataset of concepts, properties, and fine-grained relations between properties and concepts. The dataset contains 21 different semantic properties. Each property has positive as well as negative examples that allow for diagnostic experiments.

Part IV: Experiments

Part IV of the thesis presents experimental work based on the diagnostic dataset. The central goal of this part is to design experimental set-ups and frameworks for analysis that can lead to reliable insights about semantic properties in distributional representations given the methodological problems of interpretability methods. The three chapters in this part address the components of **Step 3**:

Chapter 8: Diagnostic Classification of Context-free Embeddings Chapter 8 presents two diagnostic experiments that examine context-free embeddings. A problem of many existing interpretability methods concerns the interpretation of their results. Diagnostic classification, a commonly used method for examining the information captured by latent representations, encompasses the risk of yielding misleading results. High results (i.e. results indicating that representations carry a particular aspect of linguistic information, such as a semantic property) can be caused by noise in the data or unwanted correlations rather than the target information. Chapter 8 shows how control tasks and baselines can be used as a powerful critical lens that enables a distinction between most likely valid and misleading results, in particular when combined with a methodologically informed dataset. In addition, I show how the diagnostic dataset can yield insights in an error analysis.

Chapter 9: Evidence Analysis in Two Corpora Chapter 9 presents an analysis of property-evidence in two corpora underlying the context-free models examined in Chapter 8. The chapter shows that the dataset can be used to verify results of diagnostic methods by means of corpus analysis. In addition, I examine property-evidence in the corpora with respect to the hypotheses derived from the theoretical model presented in Chapter 3. While the results are limited by small data and potential noise, it is possible to observe several initial tendencies in line with the hypotheses.

Chapter 10: Challenging Contextualized Language Models In Chapter 10, I present two approaches to the analysis of contextualized language models. I use two template-based behavioral tasks to examine pre-trained and fine-tuned models. The results of the experiments do not show clear patterns with respect property representation in language models. Rather, they indicate that the models may be exploiting other linguistic patterns captured by the models. This observation highlights the question of whether models trained on a specific task learn information that is relevant for the task or whether they exploit other regularities that accidentally lead towards high performance.

Contributions

- The research presented in Chapter 8 resulted in an elaborate set-up for diagnostic experiments involving upper and lower bounds. These bounds indicate performance ranges a diagnostic classifier should reach if it can indeed identify property information.

- The corpus analysis presented in Chapter 9 provides a verification of the diagnostic experiments.

- The corpus analysis presented in Chapter 9 shows initial tendencies about potential underlying factors that impact whether property evidence is mentioned in corpora.

- The experiments presented in Chapter 10 demonstrate a way of using the diagnostic dataset to examine pre-trained and fine-tuned contextualized language models.

- The three chapters in Part IV of the thesis examine semantic properties in distributional models from different perspectives. While it is difficult to derive clear-cut insights from individual experiments, the combination of several approaches revealed initial tendencies.

The concluding chapter contains an overview of the main findings of the research and provides an outlook for future research.

Main Findings

The research presented in this thesis has led to the following insights: Controlled diagnostic experiments in combination with the corpus analysis indicate that context-free distributional representations do not encode information about perceptual properties. For other properties, the results indicate that the representations are likely to encode fine-grained semantic categories rather than property-specific information. The analysis of contextualized representations highlighted the challenges involved in deriving reliable insights from such models. The results of the challenge task used to examine the knowledge captured by two contextualized models show that there is a considerable risk that models can exploit superficial linguistic patterns. The patterns tend to correlate with the correct answer, but do not reflect the information under investigation.

Publications, Software, and Data

The work conducted as part of the research presented in this thesis resulted in 6 publications, partly accompanied by additional resources (code and data). The publications are the results of collaborations. Table 1 provides an overview of the publications, resources, and my roles in the collaborations. This thesis is partly based on currently unpublished research. Data and code accompanying the unpublished components of this thesis are summarized in Table 2.

Chpt.	Publication	Resource
2	Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 223–233, Florence, Italy. Association for Computational Linguistics	https://github.com/cltl/semantic_space_navigation
2	Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021. Challenging distributional models with a conceptual network of philosophical terms. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2511–2522, Online. Association for Computational Linguistics *Co-supervision of Yvette Oortwijn, helped with the evaluation set-up and implementation.*	https://github.com/YOortwijn/Challenging_DMs
3	Pia Sommerauer. 2020. Why is penguin more similar to polar bear than to sea gull? analyzing conceptual knowledge in distributional models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 134–142, Online. Association for Computational Linguistics	
4	Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2019. Towards interpretable, data-derived distributional semantic representations for reasoning: A dataset of properties and concepts. In Wordnet Conference, page 85	https://github.com/cltl/semantic_property_dataset
5	Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics	https://github.com/cltl/SPT_crowd_data_analysis
8	Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286	https://github.com/cltl/semantic_space_navigation

Table 1: Overview of published work.

Chpt.	Contribution	Resource
7	Diagnostic dataset	https://github.com/PiaSommerauer/PropertyConceptRelations
8	Controlled diagnostic experiments	https://github.com/PiaSommerauer/ControlledPropertyDiagnostics
9	Corpus analysis	https://github.com/PiaSommerauer/CorpusDiagnostics
10	Masked token prediction task for contextualized language models	https://github.com/PiaSommerauer/PropertyConceptPrediction
10	Winograd-style challenge for contextualized models (*The dataset generation and experiments were implemented by Sanne Hoeken and supervised by Piek Vossen and me.*)	https://github.com/SanneHoeken/diagnostic_dataset_experiments

Table 2: Overview of currently unpublished work.

Part I

Background

The first part of this thesis presents the core concepts of this thesis, namely a perspective on word meaning through the notion of semantic properties, distributional semantics and distributional semantic models. I review existing work that considers distributional meaning representations through the lens of semantic properties and summarize what has been observed so far (Chapter 1). Subsequently, I present two case-studies in which distributional models are used to study fine-grained semantic differences (Chapter 2). These studies illustrate the current limitations of distributional models and stress the need for fine-grained explorations of what they represent.

1. Core Concepts and Related Work

1.1 Introduction

A major difficulty of investigating the meaning captured by distributional representations of words lies in their non-transparency; vector representations derived from machine learning models can act as rich sources of information for other machine-learning systems, but remain opaque to humans. While they have been shown to provide good indications of lexical similarity by means of distance in the vector space (e.g. *cat* is closer to *dog* than to *table*), it is difficult to design lexical evaluation tasks that can provide insights into specific aspects of semantic information. For example, it is difficult to assess whether the vectors also capture that cats and dogs are furry mammals, but cats tend to purr while dogs bark. This thesis aims to present a framework for investigating specific semantic properties in distributional representations. Such a framework should fulfill two requirements: (1) It should provide insights into underlying tendencies of what type of information distributional representations tend to capture. (2) The framework should consider the methodological challenges involved in interpreting nontransparent distributional representations.

In this chapter, I introduce three core concepts that constitute the main pillars of the research presented in this thesis. The explanations of the core concepts are intentionally kept 'light-weight' and accessible, as they will be taken up again and discussed in more detail in the subsequent chapters of the thesis. The remainder of this chapter is structured as follows: Firstly, I present the main assumptions behind distributional approaches to the representation of word meaning and outline the most important characteristics of distributional semantic models (Section 1.2). Secondly, I introduce the notion of semantic properties as imperfect but useful approximations of word meaning for the investigation of distributional representations and review existing semantic property datasets (Section 1.3). Thirdly, I review existing approaches to the study of semantic properties in distributional representations (Section 1.4). The chapter is concluded by an overview of common tendencies that have emerged from existing research and the limitations of existing approaches (Section 1.5) and a brief outline of the contributions of this thesis (Section 1.6).

1.2 Distributional Meaning Representations

Distributional meaning representations can provide rich semantic information but, at the same time, they are difficult to interpret for humans. What appears to be a contradiction can be understood when considering the assumptions and mechanisms behind such word representations. This section provides an outline of the theoretical ideas underlying distributional representations (Section 1.2.1), followed by an overview of traditional, context-free models (Section 1.2.2) and more recent contextualized models (Section 1.2.2).

1.2.1 The Distributional Hypothesis

Even though more recent approaches to language modeling may seem detached from linguistic theories, the core assumptions of distributional representations are grounded in the Distributional Hypothesis attributed to Firth (1957) and Harris (1954). The Distributional Hypothesis states that word meaning arises from the linguistic contexts in which a word is used. Lenci (2008) distinguishes between two possible interpretations of the hypothesis: A strong interpretation of the Distributional Hypothesis refers to a view in which linguistic co-occurrences are seen as having a causal role; they do not just merely reflect but constitute word meaning. In contrast, a weak interpretation of the hypothesis states that linguistic co-occurrence patterns can reflect word meaning but do not have a causal role. The two perspectives are not necessarily mutually exclusive; Lenci argues for a mixed view. Following this perspective, distributional co-occurrence patterns may indeed be the main source of information for abstract semantic information, such as abstract concepts, taxonomic information (e.g. the fact that a lion is an animal) and different cognitive processes. Perceptual information, in contrast, is more likely to arise from our embodied experience of the world. This type of information may, however, (at least to some degree) be reflected by distributional co-occurrence patterns. This thesis does not argue for either of the two views. Rather, this thesis aims to explore what aspects of word meaning distributional models can reflect. The fact that specific semantic information is or is not encoded in distributional representations is not necessarily evidence for the strong or weak view.

It should also be kept in mind that the distributional semantic models studied in this thesis are trained on large but finite corpora of written text. From a linguistic point of view, distributional information can refer to all aspects of language use and is certainly not limited to the patterns that arise from a specific selection of linguistic texts published on the internet and collected in a corpus. Thus, the insights from the experiments conducted in this thesis can only inform conclusions about the information encoded in specific models created on the basis of specific corpora.

1.2.2 Distributional Models

The core idea behind distributional semantic models is to represent words by generalizing over co-occurrence patterns. This principle can be implemented in a variety of ways. Traditionally, a single distributional representation captures one word form. It unites all linguistic contexts in which the word is used and can therefore not reflect different usages of polysemous words. More recently, a different approach has become popular; in this approach, words receive context-specific representations. This means that a single word form receives multiple representations for different types of contexts in which the word is used. I first explain the fundamental notions of context-free models (the former type) before outlining contextualized models (the latter type).

Context-free Models

Count-based models The simplest way of creating a distributional semantic model is to count word co-occurrences: For each word, all co-occurrences with all other words in a

corpus are recorded. The resulting counts are captured in a vocabulary-by-vocabulary matrix. Each word is represented by a vector in the matrix. Words with similar contexts will have similar vector representations. The similarity between two representations can be calculated by means of their cosine angle. When implementing such an approach, it is necessary to define how much context should be taken into account. For example, it is possible to consider the entire document in which a word appears. More common approaches define so-called context windows as a specific number of words (e.g. 2) preceding and following the target word. Context windows can also be based on syntactic information (Padó and Lapata, 2007). The nature of the context-window impacts the information represented by the model.

One of the first approaches that implemented the distributional idea of representing word meaning is called Latent Semantic Analysis (LSA Landauer and Dumais (1997)). In this approach, the context of a word is defined as the entire text document in which it appears. Each word is represented in terms of how often it appears in the documents that make up the corpus. The resulting counts are recorded in a matrix of words and documents. Each matrix column represents one document. Words with similar meaning are expected to occur in the same documents. The matrix thus also allows for measuring the similarity between documents: Each document is represented by its words. Similar documents have high lexical overlap. The size of the matrix can be reduced by means of Singular Value Decomposition.

An alternative approach to defining the linguistic context of a word is to focus on a rather narrow window surrounding the target word (e.g. two words preceding the target word and two words following it). In such an approach, each word is represented in terms of how often it co-occurs with any other word in the vocabulary within the predefined window. This approach results in a co-occurrence matrix in which each word is represented by a word vector with as many dimensions as there are words in the vocabulary. Words with similar vectors occur in similar contexts and are expected to have similar meaning. A disadvantage of purely count-based approaches is that the pure count information does not control for frequency differences between words: Highly frequent words (e.g. the determiner *the*) will have higher values than words with lower frequencies (e.g. *coffee*), regardless of their co-occurrence patterns. To account for this effect, information-theoretic count statistics based on mutual information can be used. The goal of such statistics is to provide a more accurate representation of informative co-occurrences; word pairs that occur in each other's contexts more often than independently of one another of one another should receive high values (e.g. *coffee* and *tea*).

Prediction-based models A particularly popular implementation of a distributional model relies on machine learning rather than count-based statistics. The models proposed by Mikolov et al. (2013a) and Mikolov et al. (2013b) and implemented in the Word2vec toolkit rely on a form of supervised learning. A model is trained by performing the following task: Given a particular context of an unknown word, predict the word. The model iterates over the corpus and creates increasingly accurate representations by making predictions about word-context combinations. This method is an example of supervised learning based on unlabeled data (also known as self-supervised learning). A commonly used implementation from the tool kit relies on the continuous bag of words architecture (CBOW). This architecture consists

of a shallow neural network consisting of an input or embedding layer, an intermediate or projection layer, and an output or softmax layer. The trained embedding layer is then used to represent each word.

Perhaps the most popular implementation in the toolkit relies on the Skip-Gram method. This method uses a slightly different training objective: Given a target word, the model predicts its context words. The models can be trained using different training algorithms. Traditionally, a softmax function is used to determine the prediction of the model; the function results in a probability distribution over the entire model vocabulary. The word with the highest probability is predicted. A more efficient alternative is negative sampling. This method is loosely based on the following intuition: Instead of predicting probabilities for the entire vocabulary, the model only estimates the probabilities of the target word and a small sample of 'negative' (i.e. not context) words.

The Word2vec models, while being very popular, are by no means the only successful implementations. Two other popular methods are Global Vector Representations (Glove (Pennington et al., 2014)) and fasttext (Bojanowski et al., 2017). Likely reasons for the high popularity of Word2vec are the accessibility of the toolkit as well as the large, downloadable Skip-Gram model trained on the Googlenews corpus. These representations are often used as input to neural networks trained on specific NLP tasks. Distributional representations used as input for machine learning models are generally referred to as embedding representations or embeddings.

Evaluation Typically, the quality of a distributional model is established by comparing human judgments about words to information about the same words derived from the models. Distributional models represent words in terms of their relations to other words; regardless of the underlying method, words appearing in similar contexts should receive similar vector representations and thus appear in similar areas in the distributional vector space. One criterion on which word representations can be evaluated is how well closeness and distance in the semantic space (usually measured by cosine similarity) reflect semantic similarity and relatedness. This strategy of assessing model quality is usually referred to as intrinsic evaluation.

An extensive intrinsic evaluation of different count-based and prediction-based models has been conducted by Baroni et al. (2014) and Levy and Goldberg (2014). Baroni et al. (2014) show that prediction-based models generally outperform count-based models when compared to human similarity judgments. Levy and Goldberg (2014) show that count-based models can reach equivalent performance when using a specific set of hyper-parameters.

Within NLP, distributional semantic models are mainly used as lexical representations in larger (deep) learning systems trained to perform a particular task. For many NLP problems, words constitute a highly informative source of information. Pennington et al. (2014) show that embedding representations as sole features lead to high performance for part-of-speech-tagging and named-entity recognition. Schnabel et al. (2015) provide evidence that performance on an intrinsic evaluation task does not necessarily predict performance on an extrinsic task using the example of sentiment analysis. This is to be expected, the information required for accurate sentiment prediction (mostly connotation) is not necessarily captured

by standard, intrinsic evaluation tasks. The fact that embedding representations are particularly compatible with deep learning models trained on semantic tasks indicates that they capture rich aspects of semantic information. The fact that performance on word similarity tasks does not necessarily predict performance on a semantic task may be an indication that embeddings capture aspects of semantic information that are not well reflected by overall cosine similarities, but can be extracted by means of deep learning models.

Contextualized Models

More recently, large scale contextualized language models have replaced stable word embedding representations in many NLP systems. Instead of using individual word vectors as input for supervised deep learning systems, the latest systems fine-tune entire language models on a particular task.

Language models can generally be understood as neural networks that 'read' massive amounts of text and, by doing so, acquire information about different types of linguistic or statistical regularities they observe in a corpus. In a simple, traditional set-up, a language model is trained by predicting the next word in a sequence of words. More recent implementations of language models (Devlin et al., 2019; Liu et al., 2019) are based on the transformer architecture and process text from left to right as well as from right to left. This is reflected in the acronym Bert, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2019). GPT2 (Radford et al., 2019), another commonly used language model, is a one-directional transformer model.

Generally, transformer models consist of two mechanisms: Encoding (i.e. 'reading') text and decoding (i.e. producing output). Transformers have emerged from the field of machine translation, where they are trained to encode text in a source language and produce the same text in a target language. A key component of transformer models are attention mechanisms (Vaswani et al., 2017). The purpose of attention mechanisms is to allow the model to focus on the relevant part of the input sequence for producing a particular component of the output sequence.

When transformers are used for language modeling instead of translation, transforming input text to target text is no longer necessary. Instead of transforming input text to output text, the model is trained on a masked language modeling task. In this task, words in the input sequence are masked (i.e. replaced by a [MASK] token). The model is trained on predicting the correct word for the masked token. The attention mechanisms (called self-attention in the context of an encoder-only model) allow the model to focus on words in the input sequence that are important for predicting the masked token. In addition to masked token prediction, the Bert model is also trained on predicting the next sentence. Bert's sibling model Roberta (Liu et al., 2019) is only trained on masked token prediction. Both models consist of several stacked transformer networks. Bert and Roberta only use encoding layers. GPT2 only uses decoder layers.

A main reason for the popularity of large scale contextualized language models is their suitability for being fine-tuned on a particular NLP task. Fine-tuning is done by adding a task-specific layer to the language model (e.g. a classification layer) and training it on a supervised task. During the training process, weights in the entire language model are

adjusted to 'highlight' the information that is helpful for the task the model is trained on. Contextualized models are usually evaluated in terms of their performance on NLP tasks.

It is important to realize that contextualized models differ substantially from context-free models with respect to how they represent words. Contextualized models represent words as they appear in specific contexts. Thus, the model captures information on token level, rather than on type level. In addition, contextualized models also capture representations for character sequences of words (called subwords) and can thus generalize over words it has never seen during training. Furthermore, contextualized models do not represent words as single vectors, but represent words in context in several layers of the stacked transformer network.

Nevertheless, both context-free and contextualized models rely on the same type of input data; they learn representations based on word co-occurrences taken from large text corpora. All they have access to are words in their linguistic context. Large transformer networks consist of many more parameters than simple context-free models and may thus provide richer information. Both model types, however, are founded on the Distributional Hypothesis.

1.3 Semantic Properties as a Proxy for Word Meaning

To investigate what aspects of semantic meaning distributional representations capture, it is necessary to define word meaning in such a way that it can be used for a systematic and informative analysis of distributional representations. Distributional representations have been shown to yield rich, semantic information, but they are, by no means, perfect; the best performing common sense reasoning systems still make mistakes humans would never make (e.g. Staliūnaitė and Iacobacci, 2020). Based on such findings, it seems likely that distributional data offer a partial reflection of semantic property information.

Word meaning is notoriously difficult to define. Different approaches emphasize certain aspects of word meaning, while others are left out or remain under-specified. Dictionaries capture word meaning in terms of definitions, but do not emphasize lexical relations; the computational lexicon Princeton WordNet (Miller, 1995; Fellbaum, 2010) mainly relies on lexical relations and inheritance, but under-specifies what aspects of meaning are inherited; evaluation corpora for lexical representations contain human semantic similarity and relatedness judgements, but under-specify which aspects cause the similarity or relatedness. For instance, it can be assumed that high semantic similarity indicates a certain degree of shared meaning (e.g. between the words *cat* and *dog*). However, it is not made explicit what aspects of meaning are shared.

Within psycholinguistics, approaches that emphasize the content of human conceptual knowledge about words have been proposed (McRae et al., 2005; Vinson and Vigliocco, 2008; Devereux et al., 2014). Such approaches represent words in terms of semantic properties or features that express individual components of conceptual knowledge. Rather than pre-defining properties, they have been collected from human participants by means of feature-elicitation tasks. In such tasks, human participants are presented with concepts (e.g. *strawberry*) and asked to list properties that come to mind (e.g. **red**, **juicy**, **sweet**). A major purpose of such human-elicited representations of conceptual knowledge is studying the

mechanisms behind fundamental cognitive phenomena, such as categorization, typicality, and similarity.

Just like other approaches that aim to capture word meaning, semantic properties can hardly provide a complete account of meaning. For example, it is impossible to determine how many properties are necessary to define the meaning of a concept or how fine-grained the properties should be. Nevertheless, they provide a means of analyzing individual components of meaning on the basis of empirically collected pieces of information. Analyzing distributional representations in terms of such properties could yield insights into why distributional representations do not align with human similarity judgments or why machine learning systems could not learn to perform a certain common sense reasoning task. Furthermore, understanding which semantic properties are well encoded in distributional representations could yield a more fine-grained understanding of the Distributional Hypothesis.

1.3.1 Traditional Feature Norm Sets

Semantic properties of concepts have been collected in several datasets. This section provides an overview of the most popular datasets.

McRae norms The semantic feature norm set collected by McRae et al. (2005) (usually referred to as the 'McRae norms') consists of 541 basic level concepts expressing concrete, living and non-living things. The norms have been collected from 725 participants (30 per concept). The McRae norms have been used widely in computational experiments (see Section 1.4).

CSLB norms The CSLB feature norms collected by Devereux et al. (2014, usually referred to as the 'McRae norms') constitute, to the best of my knowledge, the largest feature norm dataset. The concepts in the dataset consist of all concepts included in the McRae norms. In addition, the concepts were extended with other concrete concepts. The new concepts were selected in such a way that they are likely to share features with other concepts and thus form categories. In total, the dataset consists of 866 concepts. Features were elicited from 30 participants per concept.

Object and Action Norms The dataset collected by Vigliocco et al. (2004) consists of 456 words. In contrast to the McRae and CSLB norms, the words encompass nouns as well as verbs. The goal of the dataset was to collect features for object and action words to provide a unified representation in a single feature space.

Binder norms Binder et al. (2016) suggests 65 broad semantic properties (referred to as semantic dimensions) that can be (at least partly) justified neurologically. The datasets consists of 535 words annotated with the 65 broad properties. In contrast to the other feature norms sets, the information provided for each concept refers to an abstract semantic category rather than specific feature values. Examples of the broad semantic properties are **auditory**, **color**, and **motion**.

In general, property norm sets can be thought of as a vector space of concepts whose dimensions are defined by the total set of human-elicited features in the dataset. The values of each dimension of a concept vector consist of the number of times a particular feature was listed by a human participant. One major shortcoming of traditional feature norm sets is that they do not contain explicit negative judgments. Participants tend to list highly salient features. Consequently, the features listed for individual concepts are not necessarily exhaustive. The fact that a feature has not been listed for a concept does not mean that it does not apply to the concept. For example, the CSLB norms list the property **has_two_legs** for 16 out of 36 concepts also labeled as **is_a_bird**. 20 concepts labeled as **is_a_bird** are not annotated with the property **has_two_legs** (e.g. *duck*, *eagle*, *flamingo*). If we were to take concepts not annotated with a semantic property as negative examples of the property, these 20 concepts would constitute false negative examples of the property **has_two_legs**. Methods that aim to 'diagnose' linguistic information in latent representations, however, require reliable negative examples in addition to positive examples.

1.3.2 Augmented Feature Norm Sets

Feature norm sets have primarily been created to provide insights into psycholinguistic phenomena. Other disciplines have also been attracted to studying conceptual representations in terms of specific, empirically grounded semantic properties. Different feature norm sets have been augmented with additional information in order to use them in other experimental paradigms.

Quantified McRae Norms Herbelot and Vecchi (2016) study the phenomenon of quantification from a formal semantic perspective. To study the phenomenon with respect to empirically grounded conceptual knowledge (e.g. *some cats are black* is true, while *all cats are black* is false), they annotated the McRae norms with quantifier information. Three expert annotators (all first language speakers of English) labeled the property-concept pairs from the McRae norms with the quantifiers ALL, MOST, SOME, FEW, and NO. Herbelot and Vecchi (2016) highlight that the annotation task is not straight-forward and encompasses various semantic phenomena that can trigger disagreement among trained annotators. The resulting dataset contains explicit negative judgments. The dataset contains a substantial number of features (all features in the McRae norms), but the number of negative examples for individual properties is low. Out of 2524 features, 86 have concepts annotated with FEW or NO. Among those 86 examples, the maximum number of negative examples is 4 (for the property **is_white**).

Affordances Forbes et al. (2019) study the degree to which contextualized models can infer properties and actions afforded by properties. As part of their dataset, they extend the McRae dataset with additional concepts and a subset of human-verified negative judgments.

Discriminative attributes Krebs et al. (2018) suggest a challenging dataset consisting of concepts and discriminative attributes. Given two semantically similar concepts and an

attribute, the task is to determine whether the attribute can be used to distinguish the concepts. For instance, in the case of the concepts *airplane* and *helicoper* and the property **wings**, the property does indeed distinguish between the concepts. The dataset covers a wide variety of propertes and concepts sampled from existing dataset (e.g. the McRae norms) and added combinations of concepts and properties. Similar to the quantified McRae norms, the number of examples per property remains low.

Distributional semantic models, in particular the models in the Word2vec toolkit, have been shown to provide relatively good (albeit by no means perfect) indications of semantic similarity and relatedness. We do not know, however, if they capture information about specific semantic properties. If it is possible to distinguish distributional representations of words with respect to a particular property (e.g. **red**: *raspberry*, *blood*, *strawberry* v.s. *blueberry*, *water*, *blackcurrant*), this can indicate that the distributional representations capture information about the property.

Testing property knowledge in distributional models can yield misleading insights if the distribution of positive and negative examples does not follow a specific distribution. If the positive and negative examples fall into radically different semantic categories (e.g. **red**: *strawberry*, *raspberry*, *cherry* v.s. *chair*, *table*, *desk*) the representations can be distinguished on the basis of a large variety of properties (e.g. **is_fruit**, **does_grow**, *sweet*, **red**). In such a scenario, it is highly likely that the representations will be distinguishable, as models are known to represent general semantic (dis)similarity. If a the representations can be distinguished successfully, it is not clear whether this is due to the fact that they encode information about the target property (e.g. **red**). This risk can be mitigated by means of a controlled distribution of positive and negative examples: positive examples should only be distinguishable from negative examples on the basis of the target property. The two augmented datasets introduced in this section contain information about negative examples of properties. However, the datasets have not been designed to pose a particular challenge to distributional representations.

1.4 Diagnosing Semantic Properties in Distributional Representations

In this section, I review different approaches that have yielded insights about semantic properties in distributional models. Even though model interpretability has only gained popularity in recent years (see Belinkov and Glass, 2019), various earlier approaches have attempted to analyze the semantic content of embedding representations. The Symbol-Grounding Debate (De Vega et al., 2008) in particular, has triggered several studies that aim to investigate how much semantic information distributional models encode (e.g. Glenberg and Robertson, 2000; Andrews et al., 2009; Riordan and Jones, 2011).

I consider evidence from traditional, intrinsic evaluation methods for lexical representations (Section 1.4.1), approaches that augment distributional representations with information from other modalities (Section 1.4.2) and approaches that aim to learn mapping functions between human-elicited and distributional spaces (Section 1.4.3). I then move to approaches that

have mainly emerged from the field of model interpretability, namely diagnostic classification (Section 1.4.4) and behavioral probing tasks (Section 1.4.5).

1.4.1 Similarity, Relatedness, and Analogy

Similar words appear in similar linguistic contexts and should thus receive similar vector representations. Measuring the degree to which similarity between word vectors (measured by cosine distance) corresponds to human judgments about semantic similarity and relatedness is a core component of assessing the quality of distributional representations (e.g. Rubenstein and Goodenough, 1965; Hill et al., 2015; Bruni et al., 2014; Finkelstein et al., 2001). Context-free distributional models have been shown to correspond, at least to some degree, to human judgments (Baroni et al., 2014; Levy and Goldberg, 2014).

Semantic similarity can be interpreted as partial overlap of semantic properties. Thus, accurate reflection of similarity between two word representations (e.g. *strawberry* and *raspberry*) can give first indications that the representations may reflect semantic properties. However, a general notion of similarity (reflected by a cosine distance between 1 and 0) does not indicate which properties overlap. Partially accurate reflections of similarity do not allow for fine-grained insights about the presence or absence of certain aspects of information.

A second popular evaluation method for distributional representations are analogy tasks. Given a word pair connected by a specific semantic relation (e.g. *man* and *woman*), the task is to predict the missing component of a second pair (e.g. predict *queen* given *king*). Such analogy riddles can be solved by means of vector subtraction and addition: *king* - *man* + *woman* = *queen*. The prediction-based distributional model suggested by Mikolov et al. (2013b) has been shown to perform particularly well on such tasks. Count-based distributional models have also been shown to allow for analogical reasoning (Levy et al., 2015). This ability seems to indicate that distributional representations can capture individual components of meaning, such as maleness, femaleness or being royal.

Analogy tasks as well as analogy vector calculations have received substantial criticism. Linzen (2016) points out fundamental problems including the observation that the target vector (e.g. *queen*) in the analogy task can often be found by simply taking the vector closest to the source (e.g. *king*). Nissim et al. (2020) point out that commonly used analogy calculation methods exclude all of the three given words from the possible predictions for the missing analogy component. This limitation implies that the model has a lower chance of making a wrong prediction. In combination, these two limitations cast doubt on the assumption that simple vector addition and subtraction can be used to reason over specific semantic properties involved in analogy riddles.

Similarity, relatedness, and analogy are traditionally used to assess the intrinsic quality of context-free embeddings. Contextualized models, in contrast, tend to be evaluated by means of their performance on downstream tasks. One challenge for evaluating contextualized models with respect to the quality of their lexical knowledge is the fact that they represent words as tokens used in particular contexts. Chronis and Erk (2020) design an approach to derive type rather than token representations and show that vectors extracted from different layers of Bert perform well on standard similarity and relatedness tasks.

1.4.2 Evidence from Integrating different Modalities

As distributional models only have access to text, it can be expected that co-occurrence patterns contain little information about information that is typically perceived through other modalities (e.g. visual attributes, such as the color or shape of objects). If this is indeed the case, models that add visual information to the information encoded in text only should yield more accurate representations.

A number of early studies explore the difference in information represented by models based on text and human-elicited features. They show that representations that combined human-elicited features with distributional information tend to be more accurate than text-only or feature-only representations. The studies show that the two sources of knowledge are indeed complementary, but also encode overlapping information (Andrews et al., 2009; Silberer and Lapata, 2012; Riordan and Jones, 2011). Other approaches that directly integrate visual information with distributional models (Roller and Schulte im Walde, 2013; Silberer et al., 2013; Lazaridou et al., 2014) show similar patterns. Possible areas of information encoded in distributional data (and partly overlapping with visual data) are encyclopedic information, and function- and action-related information. Visual information, in contrast, does not seem to be well-encoded by distributional representations. A limitation of such combined approaches is that they do not test fine-grained reasoning abilities and are usually limited to qualitative evaluations or evaluations in terms of overall correlations with feature norm sets.

1.4.3 Space Alignment

Traditional feature norm datasets can be viewed as a human-created semantic space; concepts are represented by vectors whose dimensions correspond to semantic properties. A common approach used to determine whether distributional representations capture semantic features is to test whether it is possible to find a mapping between the two spaces. If distributional semantic models capture semantic information that corresponds (at least partially) to semantic features, the underlying structures of the spaces should be similar enough to align them.

Context-free models Several studies indicate that mappings from a feature norm space to a context-free distributional model space can, at least partially, be learned. Fagarasan et al. (2015) and Derby et al. (2019) experiment with different mapping functions and evaluate their approach by means of a feature prediction task: How well can the mapping predict the most important features of a word given its distributional representation? Both approaches indicate that it is possible to predict a high number of accurate features from the McRae norms (in the case of Fagarasan et al., 2015) and the CSLB norms (in the case of Derby et al., 2019). While not all predicted features necessarily correspond to features listed for the concept in the norms, the predictions are still plausible. Both studies conclude that embedding spaces are likely to capture semantic property knowledge, but might differ from human-elicited feature norm spaces in terms of what information they emphasize.

To gain deeper insights into the strengths and weaknesses of distributional representations, Utsumi (2020) conducted a mapping experiment using the neurobiologically motivated feature

norm dataset collected by Binder et al. (2016). The mapping is evaluated by means of testing how well the predicted feature vector of a word based on the distributional embedding space corresponds to the vector of the word in the feature norm dataset. The results indicate that the distributional representations provide good reflections of abstract information, but lack perceptual and spatio-temporal information.

While these mapping studies yield initial insights about a partial correspondence between the structure of human-elicited feature spaces and distributional spaces, they share the following limitations: Systematic, quantitative evaluation is limited to the features listed in the respective feature norm dataset. This means that the emphasis is almost exclusively placed on salient features. Furthermore, the experiments do not provide information about negative examples; the datasets and experimental set-ups do not allow for a comparsion to concepts that do not have a feature, as this information is not included in the feature norm sets. This limitation is particularly relevant for the study presented by Utsumi (2020), as the neurobiologically motivated norms consist of particularly broad features (e.g. **color**) and purely reflect whether a feature is relevant for a concept, but not whether it has a positive or negative association with it.

The lack of negative examples is addressed in a mapping study by Herbelot and Vecchi (2015). The study uses the McRae norms annotated with quanifier information and shows that the distributional space captures set-theoretic notions (e.g. all *tricyles* have three wheels, but only some are used for transportation). This approach indicates that, at least to some degree, the distributional vectors capture fine-grained information and may allow for fine-grained distinctions. However, it is unclear to what extent the results are caused by property-specific information or other correlations; did the model find information about the fact only some tricycles are used for transportation, or is tricycle simply dissimilar from other concepts of which most or all instances are used for transportation?

Contextualized models Some initial attempts have been made to establish to what extent pre-trained contextualized language models capture semantic information that can be mapped to human-elicited feature spaces. In contrast to context-free embeddings, contextualized models do not represent words by single vectors. Rather, they capture words as they appear in different contexts in terms of sub-words. Thus, building a vector space from a contextualized model entails a number of choices: How are word representations derived from the model? Which internal layers (or combination of layers) are used to represent the word?

Turton et al. (2020) show that a mapping can be learned between the Binder features and a space consisting of contextualized representations. To derive word representations from the contextualized model, sentences containing the target word are selected randomly. For each sentence, the representation of the target word is extracted from each layer. The representations of the target words from all sentences are averaged. The results indicate differences between the layers; overall, higher layers enable better mappings. The study indicates that a partial alignment between representations derived from contextual models and the Binder norm space can be achieved. However, it suffers from the same limitations as studies on context-free embeddings.

Abdou et al. (2021) focus on the representation of color information in contextualized models. They test to what degree a three-dimensional color space defined by lightness and two hue axes (position between red and green, and blue and yellow) can be mapped onto representations of color words derived from contextualized models. The results indicate at last a partial alignment can be achieved. Warm colors yield a better alignment with the contextualized model representations than cold colors. The study speculates that a possible reason for this finding could be that warmer colors feature more prominently in the environment (compared to cooler colors that often act as backgrounds) and therefore featured more prominently in communication. While the results indicate that there is at least partial structural correspondence between the spaces, they do not provide insights into whether color information is captured for specific concepts and whether it could be used to make fine-grained distinctions.

1.4.4 Diagnostic Classification

With the rise of end-to-end deep learning models in NLP, researchers have started to place increasingly more focus on model interpretability. If models take text as input and produce more or less accurate semantic interpretations as output, does this mean that the models capture aspects of linguistic knowledge? One of the early methodological frameworks for answering this question is diagnostic classification (Hupkes et al., 2018; Belinkov and Glass, 2019). The fundamental idea behind this approach is the following: If a latent representation (usually a layer extracted from a neural network) carries a particular piece of information, a simple classifier should learn to identify a particular piece of information if trained on a (usually relatively small) set of examples. For instance, Shi et al. (2016) used this approach to investigate whether the hidden layers of machine translation models capture different syntactic properties. The same approach has also been used to examine the content of context-free vector representations with respect to general linguistic properties (Yaghoobzadeh and Schütze, 2016), word senses (Yaghoobzadeh and Schütze, 2016) and semantic properties (Rubinstein et al., 2015).

Context-free models Diagnostic experiments on context free, distributional vectors have yielded the following insights: Rubinstein et al. (2015) conducted diagnostic classification and regression experiments using existing off-the-shelf distributional models. The models are assessed using positive and negative examples of properties extracted from the McRae norms. The results show that properties that represent a taxonomic category (such as **is_bird**, **is_-fruit**, **is_clothing**) yield comparatively high performance, while attributive properties (mostly perceptual properties such as colors and shapes) yield considerably lower performance. These tendencies hold across all four of their distributional models and across both tasks.

Diagnostic classification is particularly appealing, as it allows for targeting individual properties and in a relatively simple set-up. In contrast, learning mappings between feature and embedding spaces always relies on the structure of the entire space, and thus makes it more difficult to draw conclusions about property-specific information in embeddings.

Diagnostic classification does, however, have a number of limitations. Most importantly, it is difficult to interpret classifier performance. Perfect or close to perfect scores as well as

scores below chance level provide clear signals. However, it is unclear what scores in between these extremes mean (Hewitt and Liang, 2019; Belinkov, 2021). Connected to this problem is the fact that classifiers trained on randomly initialized embeddings can perform above chance level and even yield relatively high performance (Zhang and Bowman, 2018). This shortcoming can be addressed by using strong baselines; Hewitt and Liang (2019) suggest the use of a control task against which classifier performance can be compared. A control task should indicate how highly a classifier can perform without having access to the target information.

Contextualized models Diagnostic classification has been used to investigate a number of linguistic features in contextualized models (e.g. Jawahar et al., 2019). However, there are, to the best of my knowledge, no existing studies that use diagnostic classification to determine if layers of contextualized models capture specific semantic features. A possible reason for this lack of experiments could lie in the difficulty to interpret diagnostic experiments and their sensitivity to noise. Extracting lexical representations from contextualized models encompasses a number of choices that may introduce artifacts and cause misleading results.

1.4.5 Behavioral Experiments

Instead of studying the distributional representations directly, it is possible to study the behavior of embedding-based systems when performing a semantic task. In contrast to standard NLP tasks, such probing or challenge tasks have specifically been designed to target (or probe) specific aspects of linguistic information. Challenge datasets could be seen as a type of rigorous evaluation that aims to provide general insights about the specific linguistic abilities of a model (Lehmann et al., 1996). With increasing interest in understanding the inner-workings of deep learning models, such tasks have regained popularity in the field of interpretability (Belinkov and Glass, 2019). Challenge datasets can consist of specifically created or selected evaluation instances representative of a linguistic phenomenon. Alternatively, they can be created automatically by means of templates. For instance, Lake and Baroni (2018) created an artificial dataset to study the abilities of recurrent neural networks to generalize over the systematicity of compositional expressions.

Context-free models Several tasks have been designed to understand whether context free embedding representations can capture specific semantic information. These tasks are usually limited to using cosine distance between vectors to study specific aspects of semantic similarity or relatedness with respect to a specific semantic phenomenon.

Bruni et al. (2012) present an evaluation of text-based and image-based embedding models with a specific focus on their abilities to represent the colors of objects and distinguish between literal and metaphorical uses of color terms. Both tasks are evaluated on the basis of cosine similarity. For the object task, the cosine similarity between a noun (e.g. *crow*) and eleven color terms was measured and the rank of the correct term (e.g. **black**) recorded. This use of specifically selected evaluation data reveals general tendencies about visual information that would not be apparent from a standard test set. Visual models performed less well on

standard intrinsic evaluation sets, but outperformed text-based models on the tasks requiring visual information.

Challenge tasks have also been used to study information about afforded actions. Glenberg and Robertson (2000) use representations from an LSA model to demonstrate that distributional representations lack crucial information. In this task, a context-free distributional model is used to select a realistic continuation of a scenario. For example, given the scenario *Marissa forgot to bring her pillow on her camping trip*, the model is used to distinguish a realistic from an unrealistic continuation: (a) *As a substitute for her pillow, she filled up an old sweater with leaves.* (b) *As a substitute for her pillow, she filled up an old sweater with water.* The task is approached as follows: The sentences in the task are represented by means of averaging over the LSA vectors of each word. To determine which continuation of the scenario is realistic (i.e. afforded), the authors determine the cosine similarity between the scenario vector and the two continuation vectors. The continuation vector with the higher cosine similarity to the scenario is chosen as the correct answer. The results indicate that models can distinguish common and afforded from non-afforded situations, but are unable to distinguish between uncommon and afforded situations from non-afforded situations.

Johns and Jones (2012) present a follow-up experiment in which they compare the performance of a distributional model to a model consisting of distributional and feature norm data on the same affordance task. Rather than using the entire sentence representation, they only measure cosine similarities between the relevant words. The results of the distributional-only model confirm the observations made by Glenberg and Robertson (2000). The combined model, however, is able to achieve better performance on distinguishing uncommon but afforded from non-afforded continuations.

Contextualized models Contextualized models have yielded particularly impressive performance when being fine-tuned for a particular NLP task. Behavioral studies on contextualized models aim to study their abilities from two perspectives: (1) One goal is to investigate what kinds of linguistic regularities the pre-trained models have picked up without having been fine-tuned on a particular task. (2) Another group of approaches fine-tunes models on a particular probing task to determine what the model can learn when given training examples.

Cloze-tasks for pre-trained models A popular paradigm for the analysis of the knowledge captured by pre-trained language models are cloze tasks. Cloze tasks are sentence completion tasks that have originally been used as psycholinguistic tools that assess specific linguistic abilities. Ettinger (2020) uses a suite of psycholinguistic cloze tasks to assess different aspects of linguistic competence in pre-trained Bert models (e.g. pragmatic inferencing and understanding negation).

Sentence completion (or cloze) tasks have also been used to assess the knowledge pre-trained language models have about specific semantic properties. Weir et al. (2020) construct cloze sentences on the basis of the CSLB such as "A __ has fur." to test what they call 'human tacit assumptions'. Specifically, they test how well models can predict concepts given an increasing number of properties in the cloze sentence (e.g. "A __ has fur." eventually becomes "A __ has fur, is big, has claws, has teeth, is an animal, eats, is brown, and lives in woods." and

should be filled with *bear*). Their results provide first indications that Bert and Roberta can indeed infer the correct concepts given a combination of indicative properties. These results indicate that the models capture information about properties and concepts. However, it does not provide fine-grained insights about what properties are captured and what information is lacking.

Cloze tasks have been used to investigate the phenomenon known as the reporting bias. This bias describes the phenomenon that linguistic corpora tend to represent a distorted picture of the world, as texts (in particular new texts) tend to over-emphasize unusual and generally unexpected events (e.g. plane crashes) and under-emphasize ordinary events such as an ordinary plane journey (Gordon and Van Durme, 2013). Shwartz and Choi (2020) revisit the corpus experiments performed by Gordon and Van Durme (2013) and test the ability of language models to reflect information about people and actions. They find that language models can, to some degree, represent implied knowledge, but also tend to over-represent rare and sensational events.

Following the same intuition, Paik et al. (2021) hypothesize that people tend to omit obvious, highly implied information and predict that this type of information will thus not be captured by language models. They investigate color representation by means of a sentence completion task (e.g. "Most bananas are [MASK]."") and observe that language models are better at predicting colors of concepts whose instances can have a wide variety of colors. The finding is supported by a corpus analysis which shows that color information about concepts associated with a single color is least well represented. A complementary study was conducted by Apidianaki and Soler (2021), who target highly implied information (e.g. **red**-*strawberry*). They construct sentences reflecting quantifier information on the basis of the quantified McRae norms and find that highly implied properties are not well represented.

A limitation of the cloze tasks introduced above is that they only focus on positive examples of properties. They lack a comparison to negative examples of properties. On the basis of these results, it is hardly possible to tell whether language models can consistently distinguish between positive and (challenging) negative examples of a property.

Fine-tuning experiments While cloze-tasks can give general indications about the behavior of a contextualized language model, they do not necessarily provide an accurate reflection of the model's potential. During pre-training, all the model has to do is predicting masked tokens (and possibly next sentences). The models do not have to engage in complex reasoning. It is most likely unrealistic to expect high performance on a cloze task that requires such reasoning. It is, however, possible that the model captures enough information that could be useful when specifically trained on a task that requires reasoning.

Furthermore, sentence completion tasks usually allow for a variety of appropriate completions. The fact that a pre-trained model does not fill in the correct word does not necessarily mean that it does not have information about the semantic property in question. It merely shows that a different word received a higher probability in the masked language modeling task. The model might, however, still have the potential to learn property-concept associations when shown helpful examples in a fine-tuning set-up.

The Winograd Schema challenge (Levesque et al., 2012) constitutes a particularly illus-

trative example of what can be gained by means of fine-tuning. The challenge is supposed to constitute a particularly difficult common sense reasoning task. It consists of pronoun-resolution problems that require complex reasoning (Example 7):

(7) The *trophy* doesn't fit into the brown *suitcase* because it is too **large**. What is to large? (Possible answers: the trophy, the suitcase)

Pre-trained language models can be used to approach this task by means of masked token prediction, similar to the cloze-tasks presented above: The ambiguous pronoun *it* is masked (i.e. replaced by a [MASK] token) and the probabilities of the two candidate words (*trophy*, *suitcase*) for the masked slot are compared. The candidate with the higher probability is predicted as the correct referent. The same task can be reframed as a classification problem that then enables fine-tuning on a training dataset: Given the Winograd sentence filled with the referent candidates, predict which candidate is correct. Kocijan et al. (2019) show that fine-tuning the models on a Winograd training set can increase the performance substantially: Their fine-tuned Bert model performs about 10 accuracy points higher than the pre-trained model (0.714 compared to 0.619).

Fine-tuning a pretrained language model can have a similar motivation as diagnostic classification: If the language model captures relevant semantic information, it can learn to access this information and reason over it when given training examples. In contrast to diagnostic classification, the information is not learned based on a single vector (representing a single layer of a model), but based on the entire model with all its parameters. The parameters of the model itself are adapted during the fine-tuning process (while a vector in a diagnostic classification experiment remains frozen). The high performance gain on the Winograd Schema Challenge could be seen as an indication that the model can learn to foreground the relevant information and reason over it.

The intuition of using fine-tuning as a diagnostic task has been used to investigate the abilities of contextualized models to represent different aspects of semantic property knowledge. Forbes et al. (2019) experiment with concepts (representing objects), properties, and affordances. They fine-tune models on three tasks: Given an object, predict its properties, given an object, predict its affordances and, requiring rather complex reasoning, given an affordance, predict properties compatible with it. The results show that predicting affordances from objects can be learned successfully; models even achieve close to human performance. Predicting properties from objects works to some degree, while predicting properties from affordances cannot be achieved. The authors conclude that the models are unlikely to have seen information (i.e. indicative co-occurrences) that allow them to infer properties (e.g. being round) from affordances (i.e. rolling), as the two are unlikely to occur together. The models are too limited to make complex inferences via reasoning.

While fine-tuning can be seen as a promising diagnostic tool, it also has limitations. The fine-tuning process itself runs risk of introducing information that is unrelated to the task at hand, but happens to correlate with the distribution of correct labels. Sakaguchi et al. (2020) see reports of high performance of above 90% accuracy on the original Winograd Schema Challenge as a reason to question whether models can indeed perform the type of complex reasoning required to solve the task. They argue that the high performance is likely to be

caused by biases in the data, such as a high degree of lexical association between the trigger property and one of the candidate concepts rather than. They also consider other dataset-specific biases pointed out by Trichelair et al. (2018). Sakaguchi et al. (2020) introduce a new Winograd dataset called Winogrande. This dataset has been constructed with the goal to reduce the chance of such biases.

1.4.6 Corpus Extraction

The information captured by any language model trained on text depends on what information is expressed in the underlying corpus. If certain semantic information is not present in the corpus, it cannot be reflected by a language model. Corpus analysis can thus provide insights into the potential of language models.

Before the rise of prediction-based models, several approaches have attempted to extract semantic information from corpora in a targeted way, for instance by means of using patterns. Baroni et al. (2010) and Baroni and Lenci (2010) present a framework and model for extracting semantic property information from corpora and use it to build a vector space model. The model performs well on various semantic tasks. In particular, the model is well equipped for tasks that require information about actions and situations.

A different motivation of corpus analysis is to verify whether insights from diagnostic experiments are plausible. Paik et al. (2021) use corpus analysis to verify that the reporting bias is indeed reflected in corpus data. They find evidence of the reporting bias in three corpora and find that the same bias seems to be reflected in contextualized models (not necessarily trained on the same corpora).

Abdou et al. (2021) complement their analysis of color in contextualized models with corpus analysis. They try to identify linguistic factors in corpora that determine whether specific color information tends to be encoded in contextualized language models. They find that frequent collocations of colors in non-literal senses (e.g. *red army*) seems to correlate with low quality color representation by the language model. High diversity of syntactic roles, in contrast, seems to correlate with good color representation.

1.5 Taking Stock

In this section, I provide a summary of what is currently known about semantic property information in different word representations derived from distributional data. Semantic property information in distributional models has been investigated by means of a number of different methodological approaches. Observations range from direct extraction from corpus data to context free embedding representations and contemporary work on contextualized language models. Based on the work discussed in the previous section, I identify a number of general tendencies that seem to emerge, as well as limitations of existing approaches.

General tendencies All approaches discussed above indicate that property information is, to some degree, reflected by different distributional representations. Early approaches that investigate context-free embeddings mainly seem to find that perceptual information (in

particular color information) is not well represented by embeddings. This insight has been gained in early diagnostic classification experiments (Rubinstein et al., 2015), targeted evaluations (Bruni et al., 2012) as well as experiments that show that distributional representations can be improved considerably by means of augmenting them with perceptual information (e.g. Silberer et al., 2013; Lazaridou et al., 2014; Roller and Schulte im Walde, 2013).

Contemporary work on color representation in contextualized models shows first indications that certain aspects of color information can indeed be inferred on the basis of distributional co-occurrence patterns (Abdou et al., 2021; Paik et al., 2021; Apidianaki and Soler, 2021). A possible factor in determining whether color information is represented or not could be the reporting bias (Paik et al., 2021). Apidianaki and Soler (2021) provide first indications that highly implied information about color may remain absent from contextualized models after all.

A second tendency arising from multiple observations concerns the representation of action- and function-related properties. Rubinstein et al. (2015) show that action-related properties (as well as some aspects of encyclopedic information) yield relatively high performance in diagnostic classification experiments on context free embeddings. Results by Glenberg and Robertson (2000) and Johns and Jones (2012) indicate that context free distributional representations provide relatively accurate representations of afforded and usually performed actions. Forbes et al. (2019) provides evidence that contextualized models can reflect afforded actions successfully. Baroni et al. (2010) and Baroni and Lenci (2010) show that action- and function-related information is successfully represented by a model that relies on targeted corpus extraction.

A third aspect concerns the tendency that taxonomic information seems to be represented well by context-free embeddings based on the diagnostic experiments performed by Rubinstein et al. (2015). A possible explanation for this could be that textual data seem to be good at capturing hyponymy relations (Hearst, 1992). This tendency is in line with the hypothesis that distributional data could be the main source of abstract information concerning taxonomic categories proposed by Lenci (2008).

Limitations A central limitation of most existing approaches is that they tend to emphasize positive examples of properties and disregard the role of negative examples. In other words, focus is placed on investigating whether highly salient properties of concepts are reflected by distributional models. However, it is not investigated whether distributional representations can distinguish positive examples from negative examples of a particular property. In particular, it is not known whether distributional models can distinguish positive examples from highly similar and thus challenging negative examples. Is the information captured by distributional models fine-grained enough to determine that seagulls and penguins are birds, but penguins differ from seagulls with respect to their ability to fly? Some approaches employ datasets that contain negative examples (e.g. Herbelot and Vecchi, 2015; Forbes et al., 2019). However, they do not select positive and negative examples in such a way that they pose a particular challenge to the distributional models. Based on existing research, it cannot be said with certainty that the models capture property-specific information. High performance might simply be due to their (more or less) accurate reflection of general semantic similarity and

relatedness.

A second limitation of existing approaches is that they pay relatively little attention to the explanatory power of their experiments. Some studies try to arrive at general tendencies (e.g. Rubinstein et al., 2015; Glenberg and Robertson, 2000), but most results remain limited to showing that the language models capture semantic properties to some degree. In particular, existing studies pay relatively little attention to what the words in their datasets could reveal about general tendencies that could explain why certain aspects of conceptual knowledge are represented, while others remain absent. Contemporary studies that investigate the reporting bias (Paik et al., 2021) and highly implied information (Apidianaki and Soler, 2021) take the first steps towards identifying such tendencies.

1.6 Contributions

In this thesis, I aim to study semantic property information in distributional representations on the basis of a methodologically informed diagnostic dataset. I address the methodological limitations outlined above by specifically focusing on a dataset that challenges distributional models; the dataset should ensure that models can only succeed on diagnostic tasks if they can indeed identify information about a specific semantic property rather than by relying on other aspects of information, such as general similarity. The data used for this investigation should be indicative of possible underlying mechanisms that determine what type of semantic information distributional data can reflect. This thesis proposes a theoretical model of property expression in distributional data and a methodologically informed dataset design. I collect a diagnostic dataset that aims to fulfil the methodological requirements and enables testing hypotheses derived from the model.

The resulting diagnostic dataset is used to investigate the information captured by distributional models in the following ways: Firstly, I use the diagnostic dataset for diagnostic classification experiments on context-free embedding representations and verify the outcome by means of corpus analysis. Secondly, I use the diagnostic dataset to study contextualized models in two behavioral tasks.

1.7 Summary

This chapter presented an overview of the core concepts of this thesis and provided an overview of related research. Firstly, the chapter provided an outline of the most important assumptions underlying the Distributional Hypothesis and different distributional semantic models. Secondly, the chapter motivated the use of semantic properties for the investigation of distributional word representations and illustrated the limitations of existing property datasets when used in experimental set-ups that aim to 'diagnose' semantic properties. Thirdly, the chapter provided an overview of existing research on semantic property information captured by context-free and contextualized models and summarized the main findings and limitations. Finally, the chapter provided a brief outline of the most important contributions of this thesis.

2. Two use-cases

2.1 Introduction

In this chapter, I introduce two applications of embedding representations to the study of concepts. The use-cases illustrate the limitations of distributional models and the need for a better understanding of what aspects of meaning distributional representations reflect. The chapter summarizes two studies that explore the suitability of using context-free embedding representations for the study of specific concepts from a usage-based perspective. In particular, both studies investigate the degree to which distributional models could give accurate reflections of semantic variation and change. Both studies arose from projects that were initially not part of the research carried out for this thesis. As such, they do not focus on the central research question of this thesis.

The motivation behind both approaches is that embedding representations can provide usage-based accounts of word meaning and may reflect subtleties of meaning that are not recorded in traditional lexical resources. Both studies aim to use embedding representations for studying semantic variation and change on the basis of corpus data. Put simply, a corpus can be represented as a distributional semantic space. Two corpora can be compared against one another by comparing the structure of their semantic spaces. Semantic shifts should be reflected by changes in the position of individual vectors. Such shifts can be observed by means of comparing vector distances across the semantic spaces.

While using embedding spaces for usage-based enquiries about concepts is appealing, it is, at this point, not trivial to draw sound conclusions from such comparisons between embedding models representative of different corpora. Embedding models, in particular when trained on relatively small corpora, have been shown to be sensitive to noise and factors in the data that lead to differences between semantic spaces (Hellrich and Hahn, 2016a; Dubossarsky et al., 2017). These differences pose a risk for drawing misleading conclusions, as they are not indicative of actual semantic change. Studies that focus on shifts between specific corpora (e.g. representative of a genre, a historic period, or the writing of a specific author) tend to be comparatively small. Thus, when comparing representations from two semantic spaces to detect semantic change, there is a considerable risk of drawing conclusions based on noise, rather than meaningful change in usage.

The two studies summarized in this chapter address the challenges involved in drawing conclusions from comparatively small distributional models. These challenges are outlined in more detail in Section 2.2. The first study presents an investigation of conceptual change in the concept of RACISM over the course of the 20th century in the Corpus of Historical American English (COHA). The goal of the study is to distinguish actual semantic change from noise by means of a number of methodological checks. The summary of the study (Section 2.3) is based on the following publication:

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 223–233, Florence, Italy. Association for Computational Linguistics

The second study presents an evaluation of embedding representations created from small data for the study of philosophical concepts. The study was conducted in collaboration with a philosopher (first author of the publication), who designed a philosophical ground truth in the form of a conceptual network. We use the conceptual network to evaluate the quality of embeddings created by means of different methods specifically designed to represent words on the basis of comparatively few occurrences in a corpus. The results show that current methods are not yet good enough to enable fine-grained representations of philosophical concepts that could support the detection of subtle conceptual shifts in philosophical writing. This illustrates the need for a deeper understanding of the interaction between corpus data and embedding representations. The summary of study 2 (Section 2.4) is based on the following publication:

Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021. Challenging distributional models with a conceptual network of philosophical terms. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2511–2522, Online. Association for Computational Linguistics

2.2 Methodological Challenges of Studying Semantic Shifts

In this section, I provide a brief overview of the most important methodological problems of applying distributional semantic representations to the study of semantic variation and change. Both studies introduced in this chapter address the difficulty of distinguishing change from accidental instabilities and aim to estimating the quality and reliability of existing methods used for studying conceptual shifts, in particular with respect to comparing small corpora.

Going beyond individual words Existing studies on semantic change tend to focus on highly apparent and well-known shifts in individual words that have been documented in lexical resources (e.g. the changes in the meaning of the word *gay* from a synonym of *happy* to *homosexual*). Studying changes in more complex conceptual systems (e.g. RACISM or complex networks of philosophical terms) is less trivial. Betti and van den Berg (2014) propose the use of conceptual models to study concept change in a clearly defined and somewhat formalized way. This notion is rarely treated explicitly in applications of embedding models that aim to show conceptual shifts.

Evaluation Studies that only consider the most extreme changes (Hamilton et al., 2016a, e.g.) cannot provide accurate indications of model quality for less frequent words and subtle changes. van Aggelen et al. (2019) show that diachronic models perform considerably less well when being evaluated on a more challenging evaluation set for semantic change based on

a thesaurus. Subtle und not well documented changes, however, are of particular interest for usage-based approaches of semantic change, as such approaches have the potential of going beyond known and already documented changes. It is difficult to estimate the quality and reliability of distributional models for these types of studies, in particular when considering their potential to pick up noise rather than meaningful signals.

Sensitivity to noise A number of studies warn about the reliability of distributional semantic models for detecting change. Dubossarsky et al. (2017) illustrate that it is not known what properties in the underlying corpora are emphasized by various models and that count-based models in particular are sensitive to frequency effects. Hellrich and Hahn (2016a) point out that predictive models trained on the same data return different nearest neighbors, because they are influenced by random factors such as their initialization and the order in which examples are processed. Antoniak and Mimno (2018) present an investigation of the extent to which only small changes in the underlying corpus impact the resulting representations. They show that the impact of the processing order increases when smaller corpora are used.

Small data Diachronic general purpose corpora, such as the Corpus of Historical American English (Davies, 2002, COHA) introduced to the Computational Linguistics community by Eger and Mehler (2016), are rather limited in size. Other datasets (e.g. Google n-grams are larger, but suffer from biases (Pechenick et al., 2015) or are limited to specific genres (e.g. Google n-grams fiction) (Michel et al., 2011; Dubossarsky et al., 2015). Approaches that use embedding methods for the study of highly domain-specific data are limited to even smaller datasets. Small data have been shown to be particularly sensitive to noise (Hellrich and Hahn, 2016a).

2.3 Study 1: Studying Conceptual Change with Embeddings

In this study, we explore methods to distinguish subtle semantic change from random noise in diachronic comparisons of embedding spaces, as for instance proposed by Hamilton et al. (2016b). We approach this challenge by means of an explicit conceptual model of the well-studied, but complex concept of RACISM whose interpretation is known to have changed over the course of the 20th century. In this section, I highlight the main components and most important findings of the study[1]: I first present our explicit conceptual model that can be used to derive specific expectations about semantic change (Section 2.3.1). I then present the results of viewing apparent changes through a lens of critical methodological checks (Section 2.3.2).

2.3.1 A Conceptual Model Approach

We follow Betti and van den Berg's (2014) observation that change applies to conceptual systems. Thus, we approach the concept of RACISM as a complex conceptual system, rather

[1] Passages of the summary are taken from the original publication and have been modified to fit into the larger framework of this chapter.

than focusing on individual words or synonyms along. We distinguish four classes of words that can be relevant for studying conceptual change: (i) words referring to the core of the concept, (ii) relevant subconcepts, (iii) instances of a core or subconcepts and (iv) words referring to related concepts. We use this set-up to model the changes in the conceptual system of RACISM that have occurred over the course of the 20th century.

Conceptual model Ideally, studies of specific conceptual systems should be carried out in collaboration with domain experts. For this study, we relied on literature from various disciplines within Social Science and Humanities to select relevant words and formulate specific expectations about lexical shifts that can reflect conceptual change. We mainly rely on Barker (1981), who identifies a shift from 'old' to 'new' racism. Race used to be understood in biological terms related to visual attributes, particularly, skin color. Due to social changes (triggered by the Nazi regime's cruelties and the Civil Rights Movement), biological interpretations were relinquished as explanations for prejudice and increasingly replaced by cultural interpretations of differences between groups (Augoustinos and Every, 2007; Lentin, 2005; Morning, 2009; Omi, 2001; Wikan, 1999; Winant, 1998). We therefore identify "Culture" and "Race" as the core concepts of "Racism" investigated through the words *race* and *culture* as well as *racial* and *cultural*. The advantage of the adjectives is that they have a lower degree of polysemy than the nouns. This shifting interpretation led to different ways of defining and comparing social groups (subconcepts and instances) and different justifications for racist ideologies (related concepts) summarized in Table 2.1 and Table 2.2.

Expected changes We expect that words associated with old racism (subconcepts, instances, and related concepts) will have moved further away (i.e. the similarity of their vectors has decreased) from the core concepts. In contrast, words related to new racism should have moved closer to the core concepts (i.e. the similarity between the vectors has increased) during the 20th century. Furthermore, we expect that within the core concepts, the word *cultural* is increasingly used to describe social groups, while the biologically connotated word *racial* is avoided.

2.3.2 A Critical Consideration of Semantic Shifts

When using the approach proposed by Hamilton et al. (2016a) to study the conceptual change of RACISM in COHA and the English Google N-Grams corpus, we found several shifts that seemed to confirm our expectations. For example, the embeddings trained on the time slices of the COHA corpus indicate that several word pairs indicative of a development to new racism did indeed move closer together: *religious-racial*, *different-cultural*, *national-cultural*, *values-cultures*.

To gain insights into the reliability of these results, we stress-tested them by means of two strategies: We test whether the results hold across different model architectures trained on the same data and we test whether the changes we observed can also be observed in control words whose meaning is not expected to have changed.

Conceptual system of old racism		**target words**
Subconcepts	'Race' defined in terms of visual attributes, first and foremost skin color	*skin color* (not investigated as compound nouns are not in the model vocabularies)
Instances	Groups defined in terms of skin color	*whites*, *blacks*
Related concepts	Emphasis on a racial hierarchy	*superior*, *inferior*
	Biological justification of hierarchical structures	*genetics*
	Fear of intimacy between people of different racial groups	*marriage*, *relationship*

Table 2.1: Conceptual system and representative words of old racism.

Conceptual system of new racism		**target words**
Subconcepts	'Race' defined in terms of cultural background consisting of nationality, language and religion	*linguistic*, *national*, *religious*
Instances	Group labels of immigrants	*immigrants*, *foreigners*
	Ethnic group labels	*Jews*, *Turks*, *Arabs*
Related concepts	Emphasis on differences	*different*
	Defense of seemingly liberal values	*values*, *attitudes*, *beliefs*
	The reason for differences lies in history (rather than genetics)	*historic*

Table 2.2: Conceptual system and representative words of new racism.

Variations between models Shifts that are caused by actual change in usage should be apparent from multiple different models (of different architectures and starting with different random initializations). If the changes are stronger than seeming shifts caused by noise (i.e. random factors or meaningless frequency effects), they should arise from different models.

We test whether a subset of our initial insights are reflected from models based on different architectures. We use Hamilton et al.'s (2016a) count-based distributional semantic models, which are provided with their paper: a PPMI (Positive Pointwise Mutual Information) model and its high-density derivative SVD (Singular Value Decomposition). Though these models were less successful in detecting change according to Hamilton et al. (2016a), they reflect the

data directly without being influenced by their initialization or the order in which examples are processed Hellrich and Hahn (2016b).

We observed that some changes are only significant in a single model (e.g. *cultural-different*). For other word pairs, we observed contradictory results with significant changes in opposite directions (e.g. *cultural-inferior*). The only conclusion that remained stable and is thus supported by all models is the increasing similarity of *cultures* and *values*.

In addition to differences between models of different architectures, we also expected differences between SGNS (skip-gram with negative sampling) models trained on the same corpus but with different initializations. Models of this architecture create vector representations by iterating over the corpus and predicting whether a word-context pair is taken from the corpus or not (see Chapter 1 for a detailed explanation). They start with randomly initialized representations. Different initializations can thus lead to differences in vector representations between models trained on the same data. We trained three SGNS models for the COHA slices representative of the 1900s, 1950s and 1990s and compared the 25 nearest neighbors of *racial*. When considering the differences in the top 25 nearest neighbors of *racial* in the SGNS model trained on this comparatively small corpus, we found that as many as 14 out of 25 nearest neighbors vary among the three models trained with three different random initializations. In addition to overlaps between the neighbors, we also considered differences in neighbor rank between different initialzations of the same models. The smallest time slice of COHA with the smallest number of tokens showed the most extreme differences in rank between neighbors, while larger time slices showed less extreme variations. This finding highlights the instability of lexical neighborhoods for models trained on small corpora.

Control words Observations that hold across different models can still be a result of a bias or artefact in the data. Likewise, differences between models can also be due to noise. As an additional verification method, we test whether the changes we observed for concepts related to RACISM can also be observed for words whose usage patterns should have remained stable (henceforth control words). An illustration of the use of such control words is shown in Figure 2.1. The graphs show the changes in cosine similarity between the word pairs *races-immigrants* and *races-foreigners* (representative of the relation between race and socially defined groups) as well as the control pair *races-nurses*. In this case, the use of the control word calls the observations of change in the target word pairs into question, as the control pair shows the same patterns as the target pairs. In addition, the control word may point towards a broader semantic shift in the relation between the concepts RACE and PEOPLE. It should be considered that in this case, it is unclear whether the change relates to the biological/social or competition sense of the ambiguous word *race*. In the original study, several of the initial observations could not withstand a comparison to control words.

2.3.3 Conclusion

The study illustrates the risk of drawing conclusions about semantic changes from the comparison of embedding spaces. Based on our results, we proposed a range of recommendations for studying variation and change by means of comparing word representations across embedding

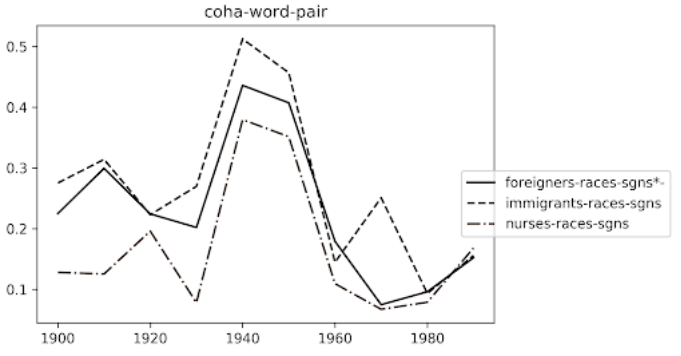


Figure 2.1: Changes in the cosine similarities between races and words representing social groups.

spaces.Our results highlight the importance of using an explicit model of the expected semantic change. Such en explicit model can be explored through multiple methods. Observed changes can be tested through a comparison to control words. Furthermore, we stress the value of comparing learning-based to count-based models, as they are sensitive to different sources of potential noise. The use of control words can serve as an additional verification. The study cannot, however, explain the different and partly contradictory tendencies revealed by models based on the same data. This illustrates the need for a better understanding of how different embedding models represent signals from distributional data.

2.4 Study 2: Evaluating Embeddings for Studying Philosophical Concepts

Study 2 constitutes an evaluation of representing philosophical concepts with different distributional semantic models specifically designed for small data. Philosophers often study how different authors used highly specialized concepts or how a single concept developed within the writing of a single author. Conclusions of such studies tend to be based on a selection of texts, rather than an exhaustive corpus, as manual close reading of the entire material is not feasible. Being able to represent and compare different collections of texts against one another by means of distributional models could thus offer a valuable tool. However, such collections of texts are comparatively small in the context of distributional models and thus prone to representing artifacts rather than meaningful signals. The evaluation conducted in this study highlights the shortcomings of distributional models for specific applications based on small, domain-specific data; while some of the approaches specifically designed for small data show promising tendencies, none of the methods yielded concept representations of sufficient quality for studying philosophical concepts. In addition, the study highlights the value of carefully constructed evaluation data created by domain experts.

In this section, I summarize the main insights obtained from the study. I briefly introduce the philosophical ground truth and how we used it for evaluating semantic representations (Section 2.4.1) and outline different approaches for representing words on the basis of small

data (Section 2.4.2). I then present selected results that illustrate the weaknesses of and open questions surrounding distributional representations (Section 2.4.3).

2.4.1 A philosophical Ground Truth

The philosophical texts under investigation in this study comprise the work of the philosopher Willard V. O. Quine, which has been digitized and processed for computational use in the QUINE corpus (Betti et al., 2020). Rather than investigating shifts within the corpus, we evaluate the quality of distributional representations created on the basis of the entire corpus by means of a conceptual network that represents central concepts and their conceptual relations in the work of Quine.

The network of concepts is based on the most important terminology defined in Quine's text *Word and Object*. The philosophical expert on the team categorized the terms into five clusters or as terms that express relations between the five clusters. Two terms in the same cluster are seen as semantically similar to each other; terms expressing relations between clusters are seen as semantically related to the terms in the respective clusters (but not necessarily similar). The entire network contains 74 clustered terms of which 43 also express relations. The clusters of philosophical terms can thus be used for similarity-based tasks, automatic clustering and classification.

An advantage of this type of highly technical ground truth is that it can be created with a high degree of agreement between experts; two independent experts reached full consensus on the accuracy of the network. While it is possible that other experts may disagree, it is remarkable that the space for interpretation left by highly technical and specialized terms is very small. By comparison, agreement reached on commonly used evaluation data for distributional models is not particularly high (e.g. Spearman Rho correlation of 0.68 for SimLex-999 (Hill et al., 2015)). As such, the expert-created evaluation set, albeit small, can be seen as a sharp tool for evaluation with a minimal risk of containing noise.

2.4.2 Methods for Dealing with Small Data

We evaluate different context-free distributional models specifically designed for small data. As a baseline model, we use a standard Word2vec skip-gram with negative sampling (sgns) model trained on Wikipedia data. We evaluate the following models against the baseline model:

Count-based model with SVD Count-based models do not encompass random factors and could thus constitute a more reliable option than models resulting from machine learning. We use a model based on mutual information (PPMI) with reduced dimensions via Singular Value Decomposition (SVD).

Word2vec with background corpus It is possible to use an existing Word2vec model as a kind of 'background' space and only train it on the target corpus (in our case QUINE) for the specific terms under investigation. The underlying idea is that the background model trained on a comparatively large corpus (we use Wikipedia data) represents a relatively accurate

semantic space into which new words based on small data can be integrated. We use two strategies for training new vectors for target words: starting with random initializations and starting with vectors that are the result of adding up representations from the context of the target word (Lazaridou et al., 2017).

Nonce2vec with background corpus Nonce2vec (Herbelot and Baroni, 2017) has been specifically designed to integrate new words based on small data into a background model. Nonce2vec also starts with vectors that result from summing words in the contexts of the target term. In addition, Nonce2vec adjusts the parameters of the model (specifically the learning rate) in such a way that the few occurrences of the target words can be maximally exploited. As with Word2vec, we experiment with two initial conditions: summed vectors (default in Nonce2vec) and randomly initialized vectors.

2.4.3 Evaluation

We use similarity, clustering and classification tasks to evaluate the different models from multiple perspectives. The motivation behind using different tasks is two-fold: (1) Testing insights by means of multiple methods leads to more reliable conclusions. (2) The different strategies might give insights into what type of approach is most promising for philosophical corpus research.

Similarity We use a simple similarity-based task using on the clusters in the evaluation set as follows: Given a target term, a term from the same cluster as the target term, and a term from a different cluster, we test whether the target term is more similar to the same cluster term or the different cluster term. This set-up allows for a simple evaluation in terms of accuracy. In this set-up, the count-based SVD model and the Nonce2vec model with additive initialization perform best. Overall, the scores remain low as the two best performing models only reach about 65% accuracy.

Cluster quality We use a measure of cluster quality (Dunn Index) to measure the coherence of the clusters defined in the ground truth in the semantic space. The score is based on the ratio between distances within clusters and cluster size. A high score indicates tight clustering and thus a high degree of separability. An accurate representation of the terms in the semantic space should reflect the clusters from the ground truth clearly and thus reach high scores. The results show that the Nonce2vec and SVD models perform best (but by no means perfectly), while the other models score very poorly.

Automatic clustering We use k-means clustering (k equals 5, i.e. the number of clusters in the ground truth) to test how well the embedding representations can be grouped on the basis of their embeddings compared to the ground truth clusters. The performance (measured by means of multiple scores) is close to random for all models with a slight advantage for Nonce2vec and SVD.

K-nearest neighbor classification As a final step, we explore how well the terms can be classified into their correct clusters by means of k-nearest neighbor classification. Again, the SVD model performs best and the best Nonce2vec model outperforms the Word2vec models. However, scores remain close to random and an inspection of the results indicate that in all models, most terms are simply classified as belonging to one of the two largest clusters.

Observation about term frequency An additional exploration of the relation between term frequency and performance on the similarity task indicates that most of the models perform better the more occurrences of a term they see. For terms with very low frequencies, the Nonce2vec models clearly outperform the Word2vec models and the SVD model. Word2vec and Nonce2vec models with additive initializations outperform their random counter parts.

2.4.4 Conclusion

The evaluation presented in this study indicates that models specifically designed to represent terms based on few occurrences yield promising tendencies. In addition, count-based models constitute a stable alternative to learning based models of promising (albeit far from perfect) quality. Despite these initial tendencies, the results indicate that distributional models, even if they are specifically designed for small data, are still far from being suitable for applications that could support philosophical research. Given the poor reflection of the clusters defined in the ground truth, it is unlikely that models could give accurate representations of meaning shifts across (potentially even smaller) corpora than the one used in this study. It should, however, also be kept in mind that the ground truth used in this study calls for highly fine-grained distinctions. Other types of term clusters with more salient differences may constitute more realistic use-cases.

2.5 Discussion

The studies introduced in this chapter constitute two use-cases in which distributional semantic representations are used directly to study concepts. The fact that distributional semantic models are trained on textual data enables a usage-based perspective that allows for observations grounded in empirical data, rather than on the basis of selective close-reading. As such, distributional models are appealing tools for research about semantic investigating shifts between corpora to study variation (e.g. differences in the writing of different authors, genre differences) and chance (e.g. semantic change over the course of multiple decades).

Both studies illustrate the difficulties of deriving reliable insights from the comparison of distributional models representative of different corpora. Study 1 highlights the variation in insights derived from multiple models based on the same corpus data. Specifically, it highlights the risk of mistaking noise for actual semantic shifts. The control tools proposed in the study can improve reliability. Nevertheless, it remains difficult to derive clear insights. Study 2 constitutes an evaluation of model accuracy for small, domain specific data. The results indicate that models specifically designed for small data outperform standard Word2vec models and that count-based models can constitute stable alternatives. However, the quality

of the word representations under investigation was still considered too low to enable the use of distributional representations as a tool for philosophical corpus research.

Both studies presented in this chapter illustrate the need for a better understanding of how distributional models based on different model architectures react to and generalize over corpus data. While the models constitute attractive tools for studying semantic phenomena from a highly empirical, usage-based perspective, it is difficult to distinguish noise from meaningful signals. Approaches that rely on the use of distributional word representations for the investigation of concepts could benefit from a better understanding how what aspects of semantic information can be represented by distributional models and how this information can be accessed.

2.6 Summary

This chapter illustrated the limitations of context-free distributional models on the basis of two studies of specific concepts in embedding models. Both studies addressed methodological challenges involved in assessing the quality and reliability of the models (Section 2.2). The first study presented a critical analysis of conceptual changes in the conceptual system of RACISM in two diachronic corpora (Section 2.3). The second study evaluated a range of distributional models for highly domain specific, specialized philosophical concepts (Section 2.4). Both studies illustrate that the interaction between different distributional methods and information reflected by data is not yet well understood.

Part II

Model

In this part, I will introduce a framework for 'diagnosing' semantic properties in distributional representations. The goal of this framework is twofold: On the one hand, the goal is to formulate testable hypotheses about the underlying dynamics that determine what type of conceptual information speakers tend to make explicit in texts and whether they do it systematically. Only information that fulfills these requirements has a chance of being represented by a distributional semantic model. In other words, one aim of the framework is to provide a usage-based, pragmatically informed account of distributional data.

On the other hand, a framework for studying distributional meaning representations has to consider the methodological constraints of analyzing and comparing such representations. Analyzing the content of distributional vectors can only be done through vector comparisons. To ensure that the comparisons can reveal something about a specific semantic property under investigation, it is necessary to select representations in such a way that the chance of discovering property-information rather than other aspects of information that happen to correlate with the property are high.

The theoretical considerations together with the methodological considerations form the basis of the design of a diagnostic dataset. I will first introduce the theoretical considerations in Chapter 3. In Chapter 4, I will outline the methodological constraints and introduce the design of the diagnostic dataset.

3. Semantic Property Information in Text

3.1 Introduction

This chapter presents a framework for the investigation of property information in distributional data. Distributional representations generalize over (usually massive) amounts of text. Thus, information that is present in the model has to arise from evidence in the texts underlying the models. It can be assumed that information has to be mentioned consistently throughout a training corpus to be reflected by distributional vectors. In particular, I present a framework of hypotheses about how linguistic evidence of semantic properties is expressed (Section 3.2) and under which circumstances it is expressed (Section 3.3). Since it is not feasible to consider concrete situations, I draw on various theoretical and empirical accounts (e.g. conversational maxims, generation of referential expressions, corpus research) to formulate hypotheses on the basis of property-concept relations. Section 3.4 presents an overview of testable hypotheses.

This chapter is based on work presented in the following publication:

> Pia Sommerauer. 2020. Why is penguin more similar to polar bear than to sea gull? analyzing conceptual knowledge in distributional models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 134–142, Online. Association for Computational Linguistics

3.2 Types of Linguistic Evidence

Before I discuss the dynamics that may determine whether a semantic property is reflected by linguistic co-occurrence patterns, I consider *how* semantic properties could be expressed in the first place. We know that co-occurrence patterns of linguistic forms in large corpora provide good indications of general semantic similarity and relatedness. Semantic similarity between word forms associated with concepts indicates that the semantic properties associated with the concepts partially overlap (Erk, 2016). Following the distributional hypothesis, word forms associated with similar meanings (i.e. similar or overlapping semantic properties) will appear in similar contexts. Thus, we expect that co-occurrence patterns can be indicative of individual semantic properties.

Co-occurrence patterns that express or point to a particular semantic property can appear in different forms and can have varying degrees of reliability. Broadly speaking, semantic properties of concepts can be expressed directly (e.g. the semantic property **red** can be expressed explicitly by the word form *red* in *red dress*) or implicitly, for instance by mentioning an activity that implies a property (e.g. *the airplane landed* implies that it was flying and can be seen as evidence for the fact that airplanes have the property **fly**). Beyond patterns that directly express a semantic property, we can also expect patterns that point to the property

indirectly via a semantic category that is associated with the property. For instance, the property **fly** is shared by most birds. Thus, co-occurrence with the word *bird* can act as indirect evidence of the property. However, it is not reliable, as not all birds fly (e.g. penguins and ostriches). I call the linguistic forms that reflect semantic properties property-evidence and distinguish different types of property-specific and non-specific evidence (Section 3.2.1 and Section 3.2.2).

Word forms appearing within a specific window of a target word form may of course be semantically unrelated to it, depending on the syntactic structure of the respective clause. In distributional semantic models that do not consider word order or even syntactic structure, this is not accounted for. Thus, words considered property-evidence are not necessarily always indicative of a semantic property and could appear in the context window by chance. Overall, though, I expect that if a semantic property is expressed systematically in the context of a target word form, the signal it provides should be stronger than this type of noise. In this section, I attempt a systematic account of different types of property evidence we can expect from co-occurrence patterns.

3.2.1 Property-specific Evidence

In this approach, I distinguish three types of property-specific evidence: Direct property evidence (Section 3.2.1), near synonyms of direct expressions (Section 3.2.1) and logical or circumstantial implications (Section 3.2.1). All three evidence types have in common that they highlight the specific property under consideration, rather than a semantic category or thematic field the property is associated with.

Direct property evidence

The most direct form of property-evidence is the word associated with the semantic property itself. While properties constitute an aspect of conceptual knowledge, they are in most cases associated with a specific word or a small group of expressions. Perceptual properties, such as colors or shapes are usually directly associated with a word (e.g. **red**: *red*, **round**: *round*). Such direct expressions of the property can appear in different morphological forms (e.g. **fly**: *flies*, *flew*, *flying*, etc.). While this type of property evidence is the most direct form of evidence, it is not necessarily always completely reliable. Linguistic forms can be ambiguous (e.g. *fly* and *flies* can be first and third person present tense forms of the verb *fly* or singular and plural forms of the noun *fly* (insect)).

Near synonyms

Words expressing a property often have near-synonyms (e.g. **cold** can be expressed directly by the word *cold*, but is also covered by the word *frozen*, the word *boiling* can be substituted for the word *hot* in certain cases and *hover* partially expresses the same activity as *fly*). Near synonyms come very close to expressing the property directly, but may also be not completely reliable indications in the case of ambiguous forms.

Implications

A third type of evidence that directly points to the target property are words that logically imply the target property or at least point to it with high probability. For instance, having a tire usually implies having a wheel, being able to give birth usually indicates female sex (but not necessarily gender). Strict logical implications are rare. Furthermore, if reduced to a simple word-co-occurrence, it is impossible to say with certainty that an individual word can be interpreted to express this type of implication. We therefore extend this evidence category beyond strictly logical implications to very likely implications.

3.2.2 Non-specific Evidence

Several expressions are likely to co-occur particularly frequently with concepts that have a property, but not point to it directly. Furthermore, such expressions can be expected to also co-occur with other concepts, but they may also occur in a wide variety of other contexts. I distinguish three types of evidence in this category: Taxonomic category evidence, thematically related words, and word associations due to cultural biases.

Taxonomic Category Evidence

A less direct source of property evidence are word forms expressing concepts that share the target property. Taxonomic categories (e.g. ANIMAL, MAMMAL, BIRD) are particularly useful devices for predicting semantic properties, as categories can be defined as collections of properties that tend to apply to their members (with varying degrees of certainty) (Rosch, 1973). Words that can indicate shared properties are co-hyponyms (e.g. *blackbird* and *robin* are co-hyponyms of *bird* and both associated with the property **fly**) and hypernyms (e.g. *bird*). Consider the following sentence taken from Wikipedia: *Many well-known birds such as hawks, eagles, kites, harriers and Old World vultures are included in this group.*[1] Such co-occurrence patterns occur frequently enough that they have been exploited successfully for knowledge-graph population (Hearst, 1992). With respect to specific semantic properties (such as **fly**), however, such taxonomic category evidence (e.g. *bird*) is only partially reliable, as several birds do not fly. Overall, co-occurrences with property-instances point to taxonomic categories that are probably associated with the property. They are not, however, property-specific and thus less reliable than property-specific evidence.

Thematically Related Words

A second source of non-specific property-evidence are expressions that are thematically related to the target property. It is likely that semantically related (but not similar) concepts (e.g. *coffee* and *cup* occur together in texts. In particular, it can be expected that thematically related concepts occur in close proximity to one another. For instance, the word *pilot* is likely to occur in situations that feature planes or helicopters and thus points towards the property **fly** for the sub-category of FLYING VEHICLES.

[1] Source: `https://en.wikipedia.org/wiki/Accipitridae` (last accessed 2021-09-06)

Associated Words due to Biases

A final group of linguistic evidence consists of words associated with a property due to cultural biases or associations that could neither be explained in terms of taxonomic relations nor situational relatedness. For example, several attributes carry a gendered interpretation and thus have the potential of pointing towards a certain gender (e.g. *beautiful*). Similarly, terrorism received much attention in the context of airport security and is associated with the property **fly**. Such words probably constitute the least reliable source of property-evidence. However, they may still be salient in co-occurrence patterns.

3.3 Expression of Linguistic Evidence

Not every mention of a concept is necessarily accompanied by the expression of property-evidence. In this section, I draw on various types of theoretical and empirical evidence to formulate hypotheses about when property evidence is likely to be expressed. Distributional models are based on large corpora consisting of texts written in a particular situation. Distributional models thus create semantic representations based on instances of usage. I draw on pragmatic tendencies and observations from corpus linguistics to formulate hypotheses about general tendencies.

When considering entire corpora underlying distributional models, it is hardly feasible to analyze all documents individually. Rather, I aim to make predictions about whether property-information is likely to be made explicit given a specific property-concept pair. In addition to the property-concept pair, I expect the property-type itself and the genre of a corpus to have an impact. With respect to text genres, I will limit the scope to two frequently used genres represented in large text corpora: Encyclopedic texts (Wikipedia) and newswire texts. In the following sections, I discuss how property-concept relations, property type, and genre may impact whether and how property information is expressed in texts. Based on theoretical and empirical accounts, I identify four phenomena which could impact the degree to which property-evidence is mentioned explicitly in texts. I count property-specific linguistic evidence as an instance in which information is being made explicit. The four phenomena generally define the relation between a property and a concept. In some cases, they also have implications for property types and the genre within which a text was produced.

3.3.1 Implied Information

A commonly used framework for understanding why speakers chose to make information explicit or not is the Cooperative Principle coined by Grice (1975). The principle consists of four maxims which aim to explain the choices speakers make when they engage in successful communication. While the maxims are primarily used to analyze conversations, written text can also be viewed as a means of communication that works within the conventions of a genre.

Semantic properties are part of our conceptual knowledge. Following the Gricean maxim of quantity, speakers only mention as much information as is required for an utterance to be understood. Thus, aspects that are already known because they are simply part of what a

concept means are unlikely to be mentioned. For example, in most cases, it is not informative to speak or write about red strawberries, because speakers can assume that this knowledge is already shared by their interlocutors. Stating it explicitly might lead to confusion rather than successful communication or serve a different communicative purpose from informing interlocutors about the common color of strawberries. Thus, in general, we can expect that implied information is not likely to be mentioned explicitly. Whether and to what degree information is implied can depend on the relation between property and concept, the property type, and the genre within which a text was produced.

Property-concept relation The degree of impliedness is not the same for all properties-concept pairs. While the color red might be highly implied for strawberries, it is probably less strongly associated with bell peppers or apples (even though both can be red). A possible mechanism underlying impliedness could be tied to property-inheritance via taxonomic categories. Red may have a higher degree of impliedness for strawberries, because it is already implied on the level of the category of RED FRUITS. The fact that a cat has four legs is most likely even more implied (and thus less likely to be made explicit, as the property of having four legs is tied to the inherited semantic category of MAMMAL). Thus, I expect that properties shared by members of a category and possibly inherited from a more general category are less likely to be made explicit.

Property type Previous research indicates that information that is already available as visual input is not mentioned explicitly. Thus, it has been proposed that information about visual properties is generally unlikely to be expressed, which has been observed in previous experiments on distributional representations (see Chapter 1).

Genre Certain text genres have the goal of capturing knowledge about the world. In particular, encyclopedic texts tend to make aspects of common sense knowledge explicit. While it is by no means feasible or helpful to explicitly mention all semantic properties of a particular concept in an encyclopedic entry, such texts can be expected to be more likely to make conceptual information explicit.

In general, most property knowledge can be considered to be implied. Thus, the 'default' assumption is that it is unlikely to be mentioned explicitly and systematically. However, certain factors (e.g. genre) may trigger explicit mentions nevertheless. Next to genre, the following three phenomena can also be expected to trigger explicit, property-specific evidence in texts.

3.3.2 Variability and Specification

Not all properties are equally strongly tied to a particular concept. Some properties can vary across instances of a concept and/or determine sub-categories (e.g. bears can be brown, black, or grey; the fur color indicates a subspecies). In some cases, there is a wide variety of possible options (e.g. cars can have almost any color).

Property-concept relations In cases in which properties can vary across instances of concepts, property-specification may indeed be necessary for informative (and ultimately successful) communication. While it is known that peppers can be red, green, or yellow, the color indicates slight differences in taste and is thus necessary information. In other cases, the property-information may be required to pick out the correct referent among various candidates (e.g. the red car rather than the white car). This notion has been used in approaches to generating referential expressions (e.g. Dale and Reiter, 1995). In such cases, property information is less strongly implied and may need to be specified.

The interaction of implied properties and variable properties has also been examined from the perspective of knowledge extraction. Gordon and Van Durme (2013) describe the discrepancy between world knowledge and what is expressed in natural language as the reporting bias: Situations that are in line with our implied conceptual knowledge do not require specific mentions, whereas unexpected or rare scenarios trigger explicit mentions. For instance, they compare the mentions to real world event frequencies of car, motorcycle and airplane crashes in proportion to miles travelled. Even though a person is far more likely to experience a motorcycle or car crash than a plane crash based on these proportions, plane crashes are reported with much higher frequency. They propose that the same tendency will hold for expected compared to unexpected properties (e.g. *a man with two legs* v.s. *a man with one leg*).

In general, variable properties often have to be specified to select the correct concept or referent when communicating. Thus, we can expect both types of property-concept relations (limited and open variability of possible properties) to trigger property evidence. Distributional semantic models, however, are more likely to reflect property-evidence if it is mentioned systematically. We can expect systematic property evidence in cases where there is a limited range of options (e.g. *bears* are either **brown**, **grey**, or **white**). In contrast, in the case of a wide (almost unlimited) range of possible properties, we are much less likely to find systematic mentions (e.g. *t-shirts* can come in a wide range of colors).

It should be noted that it is debatable whether concepts with a highly variable relation to a property should indeed be counted as examples of the property. For instance, it would be odd to consider t-shirts as examples of things which are red. What is more likely is that certain concepts have a strong association with one particular color (e.g. *dress* is strongly associated with **black**), but can also have other colors (dresses can also be green). In such situations, however, the property would be seen as a typical property of the concept, which constitutes a different property-concept relation (see Section 3.3.3). In other cases, highly variable properties might be considered rare or unusual. In the binary classification framework adopted in this thesis, the latter would be treated as negative examples. In some instances, it it might be difficult to decide whether a property should still be considered to apply to a concept. When testing a language model, such potentially ambiguous examples will be removed from the diagnostic dataset.

Property-type A group of properties that tends to be particularly variable are color properties. In cases in which color is variable, it may indeed be likely to be mentioned. In cases in which there are many options, however, the mentions may not be systematic enough to arise

from distributional data. The variability of color properties is a potential exception to the general expectation that visual properties are not likely to be mentioned explicitly. Explicit property mentions due to variability is particularly likely to affect a subset of property-concept pairs with color properties.

3.3.3 Property-Illustrations

Highly implied property knowledge can be expressed explicitly to fulfil a communicative goal beyond the expression of necessary distinctions. Corpus research by Veale and Hao (2007) and Veale (2013) shows that concepts that are so strongly associated with a property that they can serve as illustrations of it can be extracted from corpora. For instance, particular colors can be described in terms of particularly good examples that exhibit them (e.g. *as red as blood*, *as black as ebony*).

Property-concept relations Concepts that tend to be used for property illustrations tend to be typical examples of a property. This type of typicality relation is likely to correlate with a particularly strong association between a property and a concept. Association strength by itself, however, is most likely not sufficient to trigger explicit mentions. For instance, properties listed by many participants for a concept in property norm datasets show strong associations, but do not necessarily only capture pairs in which the concept can be used to illustrate the property (e.g. **green**-*broccoli*). In many cases, strong association may indicate that the property is typical of the concept, rather than the other way around.

It is questionable whether such illustrative expressions occur systematically enough to be reflected by distributional representations. If concepts can indeed be used to illustrate a property, they should at least have higher chances for more explicit property mentions than concepts that are linked to the property merely via strong association.

Genre Property illustrations may be more likely in texts that have a higher degree of imagery (such as different literary genres). Thus, it can be expected that they are not particularly common in encyclopedic texts. News texts are not particularly well known for imagery either, but corpora consisting of a wide selection of news texts have higher chances of containing slightly more creative writing than encyclopedic texts.

3.3.4 Affordance

Within cognitive linguistics, afforded actions form a central component of semantic knowledge (Gibson, 1954; Glenberg, 1997). Afforded actions refer to the actions available to a person in a specific situation. For instance, a candle can be lit or extinguished. In many cases, a candle has a round shape and could also be rolled across a table. Such actions can be seen as building blocks of bigger events. Thus, it is likely that they are also mentioned in natural language.

Property-concept relations Existing research has shown that distributional representations reflect afforded actions (Fulda et al., 2017), but they tend to fail at distinguishing unusual but possible from impossible actions (Glenberg and Robertson, 2000). While it is likely that

commonly performed activities will be mentioned systematically (e.g. lighting a candle), it is unlikely to find much evidence for unusual activities (e.g. rolling a candle).

Actions or uses can also provide indications about other properties of a concept (e.g. attributes or parts). For instance, things that are round tend to roll (e.g. a bowling ball), things that are used for cutting tend to have a sharp edge (e.g. a knife). Explicitly mentioned activities can thus be a reflection of other properties by means of implication. It can be expected that properties that enable common uses or activities are reflected by means of property-evidence in the form of implications. If such activities or uses are common, they have high chances of being mentioned systematically.

Genre Genres can be expected to differ with respect to their emphasis on events. It can be expected that news texts are centered around events, while encyclopedic texts tend to emphasize the conceptual level. Thus, it can be expected that afforded actions and affording properties are better reflected in news texts than in encyclopedic texts.

3.4 A Framework for Testing Hypotheses

In this section, I present hypotheses that can be tested based on the considerations about property evidence and factors that may trigger evidence expressions. I first focus on semantic relations and then consider possible interactions with genre and property types.

Semantic Relations

Table 3.1 presents specific semantic relations based on the factors presented in the previous section. A combination of a property and a concept can be defined by one or multiple relations. The examples presented in the table should be read as illustrations of a relation.

Each semantic relation is tied to a hypothesis about the type of property evidence expected to be found in corpora (summarized in Table 3.2). The table also shows a mapping from relations to quantifier information similar to the Quantified McRae norms (Herbelot and Vecchi, 2016). For instance, the variability relations express that a property applies to a subset of instances, rather than all instances of a concept. This information is relevant for contrasting positive and negative examples of a property.

The hypotheses about property-evidence should be interpreted as follows: A particular property-concept pair (e.g. **green**-*broccoli*) can be characterized by multiple semantic relations (e.g. `implied_category`, `typical_of_concept`). If none of the relations are expected to trigger evidence, we do not expect to find evidence in the distributional data. If one of the relations is expected to trigger property evidence, we expect this relation to 'override' the other relations. For example, the pair **sweet**-*sugar* can be described by `typical_of_concept`, `typical_of_property`, and `affording_activity`. In this case, the relations expected to trigger with property-evidence override the relation `typical_of_concept`.

The summary presented in Table 3.2 allows for testing hypotheses about the amount and type of property-evidence found in distributional data. I derive general and phenomenon-

factor	relation	explanation	example
impliedness	`implied_category`	The property is strongly implied because it is part of the knowledge about the semantic category of the concept.	**mammal**-*cat*
variability and specification	`variability_limited`	Instances of the concept may vary with respect to the property. There is only a limited range of options of property values of the property-type.	**yellow** - *bell pepper*
	`variability_open`	Instances of the concept may vary with respect to the property. There is a wide, almost unlimited range of options of property values of the property-type.	**red**-*car*
typicality	`typical_of_property`	The concept is very strongly associated with the property and serves as an illustration of it.	**red**-*blood*
	`typical_of_concept`	The property is very strongly associated with the property. It is one of the first properties that come to mind when thinking of the concept.	**green**-*broccoli*
affordedness	`afforded_usual`	The property expresses an action or use of instances of the concept.	**cut** - *scissors*
	`afforded_unusual`	The property expresses an action or use of instances of a concept that is possible, but not common or frequent.	**clean glasses** - *t-shirt*
	`affording_activity`	The property enables a use or action of instances of the concept.	**round**-*bowling ball*

Table 3.1: Property-concept relations with explanations and examples.

.

specific predictions, and predictions about the three different types of subsets identified by the relations (ALL, SOME, FEW-NONE):

- **General**: Evidence representation should be stronger for pairs characterized by relations hypothesized to trigger systematic evidence expression than relations not hypothesized to trigger systematic evidence expression.
  - Positive relations hypothesized to trigger evidence:
    * `typical_of_property`
    * `affording_activity`
    * `afforded_usual`
    * `variability_limited`
  - Positive relations hypothesized not to trigger evidence:
    * `implied_category`
    * `typical_of_concept`
    * `afforded_usual`
    * `variability_open`
  - Negative relations:
    * `rare`
    * `unusual`
    * `impossible`
- **Phenomenon-specific**: While the overall tendency described in the general hypothesis may hold, differences between individual relations may not necessarily follow the separation outlined above. Therefore, I formulate phenomenon-specific hypotheses:
  - *Impliedness*: Pairs only characterized by `implied_category` should have a lower degree of property-evidence in the context of the concepts than pairs characterized by any other positive relation.
  - *Property-illustration and typicality*: Pairs characterized by `typical_of_property` should have a higher degree of property-evidence than pairs characterized by `typical_of_concept`.
  - *Affordedness*:
    * Pairs characterized by `affording_activity` should have a higher degree of property evidence than pairs characterized by `afforded_unusual`.
    * Pairs characterized by `afforded_usual` should have a higher degree of property evidence than pairs characterized by `afforded_unusual`.
  - *Variability*: Pairs characterized by `variability_limited` should have a higher degree of property evidence than pairs characterized by `variability_open`.
  - *Negative relations*:

    - Pairs characterized by `rare` should have a higher degree of property-evidence than pairs characterized by `unusual` and `impossible`.
    - Pairs characterized by `unusual` should have a higher degree of property-evidence than pairs characterized by `impossible`.

- **Subsets**:
  - Pairs in the ALL category should have a higher degree of evidence than pairs in the SOME and FEW-NONE categories.
  - Pairs in the SOME category should have a higher degree of evidence than pairs in the FEW-NONE category.

Property-type

Based on the considerations with respect to property-type, I expect the following tendencies:

- Most pairs that involve a property expressing an action or use will be characterized by the relation `afforded_usual`. Thus, it is likely that most concepts associated with these properties will exhibit property-specific evidence.
- Pairs involving properties that are closely tied to taxonomic categories (e.g. **lay_eggs**-*seagull*) are likely to be characterized by the relation `implied_category`. They are unlikely to be characterized in terms of any of the `variability` relations. Unless the properties are also tied to specific activities or uses, concepts involved in such pairs are unlikely to exhibit property-specific evidence.
- In general, property-evidence for color properties is expected to be low. However, for concept-property pairs characterized by `variability_limited`, property-evidence is expected.

3.4.1 Genre

Based on the considerations with respect to genre, I expect the following tendencies:

- Concepts involved in pairs characterized by the relation `implied_category` may exhibit more property evidence in encyclopedic texts than in news texts.
- Concepts involved in pairs characterized by action- and use-related properties may exhibit more evidence in news texts than in encyclopedic texts.

3.5 Summary

In this chapter, I have presented a framework of hypotheses for the analysis of property-evidence in text. The core of the framework consists of semantic relations that characterize property-concept pairs. In addition to these relations, I considered the impact of and interaction with property types and characteristics of genres. The focus has been placed on property-specific evidence in corpus data.

In the following chapter, I will consider fundamental methodological considerations that have to be considered when testing or 'diagnosing' semantic property evidence in distributional representations. Based on these considerations, I will present a design for a diagnostic dataset that (1) can be used to test the hypotheses presented in the current chapter and (2) adheres to the methodological constraints imposed by distributional representations. In this context, non-specific property evidence will receive more attention.

set of instances	factor	relation	evidence
most - all	impliedness	`implied_category`	non-specific
	typicality	`typical_of_concept`	non-specific
		`typical_of_property`	prop. specific + non-specific
		`affording_activity`	implications + non-specific
	affordedness	`afforded_usual`	prop-specific + non-specific
		`afforded_unusual`	sparse to none
some	variability	`variability_limited`	prop-specific
		`variability_open`	sparse
few-none	negative cases	`rare`	sparse
		`unusual`	sparse - none
		`impossible`	none
creative		`creative`	sparse - none

Table 3.2: Summary of linguistic hypotheses about semantic relations and types of evidence. The semantic relations are sorted with respect to the subset of concept instances they apply to.

.

4. Methodological Framework and Dataset Architecture

4.1 Introduction

The fundamental problem of studying distributional representations (whether they are taken from a context-free or from a contextualized model) is that they represent information through structure. That is, they can only be informative when put into relation to other representations. The representation of the word *penguin* in isolation cannot be interpreted. The vector only receives meaning when considered in relation to other word vectors in the semantic space. The meaning of individual words is defined by similarities to and differences from one another. For example, it can be expected that the representation of *penguin* is close to the words *bird*, *animal*, and *polar bear*. Moreover, *penguin* is probably closer to *bird* and *animal* than to *table* and *lamp*.

Comparisons on the basis of distances in the semantic space can provide indications about the quality of embedding vectors. However, they are limited by the fact that we cannot know what causes their closeness or distance. For example, it is likely that the representation of *penguin* is close to *bird* and *polar bear*. The semantic relations between *penguin* and *bird* is different from the relation between *penguin* and *polar bear*. Consequently, the semantic properties shared by *penguin* and *bird* are different from the semantic properties shared by *penguin* and *polar bear*. These differences do not arise from a comparison of vector distances in the semantic space. This thesis aims to 'diagnose' semantic properties through testing whether word representations can be distinguished on the basis of a specific semantic property. For example, I test whether it is possible to perform the following task on the basis of distributional representations (Example 8):

(8) Which of the following concepts can be described by **can fly**?
   a. *seagull*
   b. *table*
   c. *airplane*
   d. *penguin*
   e. *bee*
   f. *strawberry*

Informative diagnostic tasks crucially depend on an informative dataset. This chapter presents a number of methodological considerations about the analysis of distributional representations that inform the design of such a diagnostic dataset. A dataset that can reveal whether distributional vectors contain property-specific information in diagnostic experiments

has to adhere to a number of constraints (outlined in Section 4.2). Based on these constraints, I present the architecture of the diagnostic dataset: Section 4.3 outlines the selection of properties and Section 4.4 describes that selection of example concepts. Section 4.5 presents an analysis of the candidate properties and concepts.[1]

The properties and candidate concepts presented in this chapter constitute the architecture of the dataset. To make the dataset informative with respect to the hypotheses presented in the previous chapter (Chapter 3), the resulting property-concept pairs still have to be labeled with property-concept relations. The subsequent part will outline the annotation process and an analysis of the finished dataset.

This chapter is based on the following publications:

> Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2019. Towards interpretable, data-derived distributional semantic representations for reasoning: A dataset of properties and concepts. In Wordnet Conference, page 85
>
> Pia Sommerauer. 2020. Why is penguin more similar to polar bear than to sea gull? analyzing conceptual knowledge in distributional models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 134–142, Online. Association for Computational Linguistics

A first version of the methodological considerations and a small pilot dataset were presented in the following publication:

- Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286

4.2 Detecting Properties in Distributional Vectors

In this section, I present the core assumptions behind the interpretability methods I use to investigate whether distributional models represent semantic properties. Initially, the dataset has been designed for the analysis of context-free distributional representations. As context-free models and contextualized models have different properties, I design different diagnostic tasks for each model type (see Chapter 1 for a description of context-free and contextualized models). Context-free representations have gained popularity as they work particularly well on downstream tasks in which they are used as input for machine learning models (usually neural networks) trained on a specific task (e.g. Socher et al., 2013). This indicates that context free vectors capture information that can be exploited successfully by neural classifiers. Therefore, I employ diagnostic classification to analyze the content for context free vectors (Section 4.2.1).

[1]The code used to compile the candidate dataset can be found at `https://github.com/cltl/semantic_property_dataset`

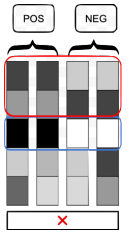
Contextualized language models have been shown to perform particularly well when fine-tuned on a specific task (e.g. Devlin et al., 2019). While it is possible to extract vector representations of individual words from contextualized models and use them in a diagnostic classification task, this set-up comes with a number of additional choices and possible parameters which complicate the analysis: Firstly, contextualized models capture words in context and deriving representations of individual words requires additional steps and may introduce noise (Rogers et al., 2020). Secondly, contextualized models capture words in context on multiple layers. In a fine-tuning set-up, the model can exploit information from all layers. In contrast, in a diagnostic classification set-up, it would be necessary to either use individual layers or concatenate them. Neither option realistically represents the information the model has at its disposal when used for language modeling or a specific task. Therefore, I opt for a challenge task designed to reveal property knowledge (Section 4.2.2). Both paradigms impose methodological constraints that have to be considered in the dataset design (Section 4.2.3).

4.2.1 Diagnostic Classification

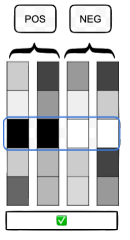
The fundamental assumption behind diagnostic classification is the following: If information is present in a latent vector representation, a classifier should be able to learn to recognize this information and classify unseen examples correctly. Translated to semantic properties, this means the following: If a distributional vector representation of a word carries property-evidence (e.g. **being able to fly**), a binary classifier should be able to learn how to distinguish positive examples (e.g. *seagull*) of the property from negative examples (e.g. *penguin*). The property-concept dataset is primarily designed to be informative in such a diagnostic classification set-up. The details of the experimental set-up are introduced in Chapter 8.

A fundamental problem of this approach is that successful classification does not necessarily mean that the classifier did indeed identify property-specific information. Consider the following example: Suppose a classifier is trained to distinguish positive examples of the property **red** from negative examples. Positive examples of the property could, for instance, be words referring to red fruits: *strawberry*, *raspberry*, *cherry*, *pomegranate*. Negative examples could consist of words referring to green garden plants (e.g. *ivy*, *willow*, *privet*, *beech*). If the classifier successfully learns to distinguish these examples, it could have identified information about a number of semantic aspects that allow for a correct distinction: It could have detected information about the fact that all positive examples tend to be used as food, that they tend to have a sweet taste, that they are juicy or that they are often used in combination with each other. Vice-versa, it could have learned that the negative examples tend to be categorized as shrubs, are offered for sale in garden centers and generally have nothing to do with food. Whether the classifier has identified information about the target property **red** or any of the other semantic aspects that could lead to successful classification remains an open question.

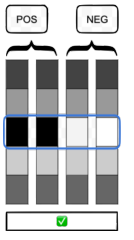
The simplified example illustrates how the distribution of positive and negative examples determines what a diagnostic classification experiment can reveal. Any supervised classification approach relies on finding regularities that are shared by all (or most) positive examples. Negative examples do not share these regularities. Ideally, the only shared feature



(a) Target property (present in black, absent in white) and other dimensions correlates with positive and negative classes.



(b) The only dimension correlating with the positive and negative classes is the target dimension of the target property.



(c) Highly similar positive and negative examples that can only be distinguished by the dimension of the target property

Figure 4.1: A schematic representation of vectors of positive and negative examples of a property. To ensure that shared and distinguishing patterns identified by a classifier are representative of the target property, positive and negative examples should only be separable based on the target property.

of all positive examples should be the semantic property under investigation. However, it is possible that other features also allow for distinguishing positive from negative examples. In such a scenario, the dataset offers multiple routes to the correct classification output. High performance does not necessarily mean that the classifier detected the property in question. In contrast, if high performance can only be achieved by detecting the target property, high performance indeed an indication that the embedding representations carry information about the target property.

The intuition behind such a dataset distribution is illustrated in Figure 4.1. If the positive examples can be distinguished from the negative examples by means of multiple aspects of semantic information (for instance, because they share the same category) successful classification does not indicate property information (Figure 4.1a). In contrast, high classifier performance is indicative if the property information is the only aspect that positive examples have in common and negative examples do not (Figure 4.1b) or if the property information is the only aspect that distinguishes positive from negative examples (Figure 4.1c). The latter two scenarios can be achieved by either a highly diverse set of positive and negative examples of the property (e.g. taken from many different semantic categories) or a distribution in which positive examples are highly similar to negative examples.

4.2.2 Challenge Task

To analyze contextualized language models, I design a task in which a fine-tuned contextualized model has to apply property knowledge in order to perform well. Property knowledge encompasses aspects of lexical- as well as common sense knowledge. An existing framework for assessing this type of knowledge has been proposed in the form of the Winograd Schema Challenge (Levesque et al., 2012). The challenge consists of pronoun resolution problems that can only be solved by reasoning various semantic aspects expressed by individual words or phrases. An example of a Winograd pronoun problem is shown below (Example 9):

(9) The *trophy* doesn't fit into the brown *suitcase* because it is too **large**.

To resolve the pronoun *it* successfully, a model has to engage in complex reasoning about the fact that if something does not fit into a container, it means that the thing (i.e. the trophy) is too large (rather than the container).

The same framework can be used to design a task that specifically targets semantic property knowledge (Example 10):

(10) Yesterday, I used a *knife* to slice an *orange*. Unfortunately, it was so **juicy** that I stained my t-shirt.

To resolve the pronoun correctly, a model has to know that oranges are more likely to be juicy than knives. Resolving the task involves the following steps: (1) The model has to recognize the aspect that links to pronoun to one of the candidate concepts (*orange* or *knife*). This trigger is expressed by the adjective *juicy*. (2) The model has to decide which candidate is more likely to be described by *juicy*. The task has been adapted to be more suitable for contextualized language modeling by recasting it as a 'fill in the blank' type of task (Sakaguchi et al., 2020). Example 10 is then transformed to Example 11:

(11) Yesterday, I used a *knife* to slice an *orange*. Unfortunately, the __ was so **juicy** that I stained my t-shirt.

The use of positive and negative examples from the diagnostic dataset can be used to create Winograd-style examples. Chapter 10 presents a template approach for automatically generating a large number of such examples and an evaluation of contextualized models on the generated Winograd-style property dataset.

How should a language model be able to perform such a complex task? The underlying idea is the following (refer to Chapter 1 for a detailed explanation): The model learns semantic information by means of performing a language modelling task (in the case of Bert, this would be masked token prediction and next sentence prediction). In the fine-tuning phase, the model is trained on a supervised classification task: Given the two possible versions of the sentence pair, (either completed with *orange* or *knife*,) predict the correct one. If the model has captured the relevant information about the candidate concepts and the trigger word in the pre-training phase, the supervised fine-tuning process should guide the model towards exploiting this information for predicting the correct version of the second sentence.

The danger of this approach is that the model can rely on a superficial word association between the correct concept and the trigger word (i.e. the word expressing the semantic property), as pointed out by Sakaguchi et al. (2020). This danger can be mitigated by means of selecting candidate concepts in such a way that superficial associations are not helpful. For instance, two similar concepts from the same semantic category (e.g. *penguin* v.s. *seagull*) are more difficult to distinguish than two concepts from entirely different categories (e.g. *orange* v.s. *knife*). To distinguish *orange* from *knife* with respect to **juicy**, a model might simply make the correct choice because of a superficial association between *orange* and *juice* or *juicy*, which is simply not there for *knife* and *juicy*. If a model does indeed have an association between *seagull* and **fly**, but not *penguin* and **fly**, it is more likely to be a reflection fine-grained property knowledge.

4.2.3 Dataset Requirements

Both approaches described above impose a number of requirements on an informative diagnostic dataset of properties and concepts.

Verified Negative Examples

Any diagnostic approach that relies on a comparison between positive and negative examples relies on the assumption that the positive and negative examples are indeed correct. A common pitfall of existing approaches is to extract both positive and negative examples from feature norm datasets (see Chapter 1). As a basis for the selection, it is common to use the feature production frequencies; concepts for which a property has been listed multiple times are taken as positive examples, whereas concepts for which a property has not been listed are taken as negative examples. This approach works well for positive examples, but runs risk of resulting in a relatively high number of false negatives. For instance, in the CSLB norms, 36 concepts are labeled as **is_bird**, but 20 out of those 36 concepts are not labeled as **has_two_legs** (e.g. *duck*, *eagle*, *flamingo*). Thus, taking all concepts for which the property **has_two_legs** has not been listed as negative examples of the property would result in at least 20 false negatives.

Sufficient Positive and Negative Examples

A second requirement for any approach that relies on supervised learning is a sufficient number of examples for training and testing. The assumption behind diagnostic classification is that a classifier should be able to learn information if it is encoded based on a small set of examples. Still, the number of examples has to allow for a test set that is big enough to draw meaningful conclusions from the performance. In the case of semantic properties, it has to be considered that not all examples may carry property-information (see Chapter 3). Thus, it has to be anticipated that a certain proportion of examples may simply act as noise. To ensure that the diagnostic approach is still robust, the dataset size per property should be large enough to allow for this type of noise.

In addition to the constraints imposed by machine learning, the datasets should also contain a selection of different properties and enable comparisons between properties of different property-types (e.g. perceptual properties, activities and functions) Even though the hypotheses I introduced in Chapter 3 rely on the property-concept relation, previous research indicate that certain types of properties are better represented by distributional data than others. For example, Rubinstein et al. (2015) find that embeddings provide information about encyclopedic properties, but not about perceptual properties (refer to Chapter 1 for details). A sensible comparison to previous findings thus requires conclusions about individual properties.

Diverse Examples

A particular risk of diagnostic classification is that the classifier learns a distinguishing feature that happens to correlate with the positive examples of a property, rather than property-specific information. As illustrated by an example involving the property **red** in Section 4.2.1, it may

learn to identify the category of fruits, rather than words with the property **red**. To avoid such misleading conclusions, the positive and negative examples of a property should be as diverse as possible. This diverse distribution should create a situation in which the property-specific information is the only aspect that connects positive examples and distinguishes them from negative ones. This requires (a) a selection of semantic properties that apply to a diverse set of concepts and (b) a concept selection process that targets a wide range of concepts.

High Similarity between Positive and Negative Examples

In both diagnostic approaches, positive and negative examples with a low similarity increase the risk of misleading results: In the case of diagnostic classification, a low overall similarity between positive and negative examples may result in a situation in which a classifier performs highly based on learning multiple categories associated with the target property, rather than the target property.

In the case of the Winograd-style challenge, low semantic similarity between positive and negative examples also poses a risk. In such cases, it is possible that the positive example has closer lexical association with the property than the negative example. For instance, given the property **yellow**, the positive example *apple* and the negative example *idea*, the correct referent can be picked based on the fact that concrete things can have colors, but abstract concepts usually do not. This does not reveal whether a model 'knows' that apples can be yellow. Thus, a combination of highly similar concepts that can only be distinguished by means of the target property provides stronger indications that the model can capture fine-grained semantic properties rather than exploit correlations to perform the task. A diverse set of examples can lead to more robust results, as high performance on a diverse set of positive and negative examples reduces the chance that the model exploited accidental correlations.

4.3 Selection of Properties

This section presents the semantic properties selected for the diagnostic dataset. The properties were selected manually from the properties represented in the CSLB norms based on the following two rationales:

1. It should be possible to find a large set of words that have the property.
2. The property should apply to concepts from many different traditional taxonomic categories to increase the diversity of positive examples.

In addition to these two rationales, I selected different property types to enable a comparison to previous findings. Various studies provide indications that visual information tends to be absent from distributional representations, while information that is strongly tied to taxonomic categories is represented well. I specifically include these property types to test whether previous findings hold given the example distribution outlined above.

Table 4.1 shows the 21 selected properties with their property types. The types are based on the categorization used in the CSLB norms. Modifications have been made for the following reasons:

- Different types of perceptual properties may have a different relevance for afforded action (e.g. uses) and should be treated separately.
- I treat part properties as feature type rather than as visual perceptual properties, as they tend to serve as distinguishing features between taxonomic categories (Miller, 1995) and are thus closely tied to specific taxonomic categories. They are much closer related to taxonomic properties than visual-perceptual properties.
- I treat properties that can be seen as a combination of multiple ontological properties and depend on interpretation as 'complex' properties (e.g. multiple properties in combination lead to the fact that tigers are interpreted as dangerous animals).

type	**type (CSLB)**	**properties**
taxonomic	taxonomic	lay_eggs
part	visual-perceptual	wheels, wings
function/action	functional	roll, fly, swim
complex	encyclopedic	dangerous
complex	functional	used_in_cooking
perceptual (taste)	other perceptual	juicy, sweet
perceptual (temperature)	other perceptual	cold, hot, warm
perceptual (color)	visual-perceptual	black, blue, green, red, yellow
perceptual (shape)	visual-perceptual	round, square
perceptual (material)	visual-perceptual	made_of_wood

Table 4.1: Properties selected for the diagnostic dataset.

4.4 Selection of Concepts

In this section, I describe the selection of candidate concepts for each of the properties. It is important to note that this step only concerns the selection of example candidates. The candidate concepts still need to be verified and annotated with the semantic relations introduced in the previous chapter (Chapter 3). The annotation task and process are outlined in Chapter 5 and Chapter 6.

Ideally, each property should have a balanced set of positive and negative examples whose distribution follows the requirements outlined in Section 4.2.3. To achieve this, I exploit existing resources (Section 4.4.1) and a distributional model (Section 4.4.2). For most properties, these two steps result in a large number of candidate concepts. I use a sub-sampling method to select candidates with respect to a number of linguistic features (Section 4.4.3).

4.4.1 Extraction from Existing Resources

I extract positive and negative example candidates from feature norm datasets, lexicons, and a stereotype dataset via different search strategies. These resources primarily provide information about positive examples, but also allow for the extraction of good negative

type	resources
feature norm sets	McRae norms (McRae et al., 2005), CSLB norms (Devereux et al., 2014)
lexicon	WordNet (Fellbaum, 2010; Miller, 1995) ConceptNet (Speer and Havasi, 2012)
stereotype data	concepts representing stereotypes of properties (Veale, 2013)
feature norms negative extension	subset annotated on top of the CSLB norms Sommerauer and Fokkens (2018), quantified McRae norms (Herbelot and Vecchi, 2015)

Table 4.2: Overview of resources used for finding positive and negative property candidates.

candidates. In addition, small sets of negative examples were extracted from the quantified McRae norms (Herbelot and Vecchi, 2016). Another small set of negative examples was verified by hand in a pilot study. An overview of all resources is provided in Table 4.4.1. In this section, I outline how the different resources were exploited to select suitable candidates.

Direct Property Searches

Directly searching for a property in the feature norm sets returns a set of reliable candidates for positive examples. The lexical resource ConceptNet (Speer and Havasi, 2012; Speer et al., 2017) also allows for such direct searches. It links concepts to different attributes via semantic relations, such as *HasProperty*. ConceptNet also records negative associations: The relation *NotHasProperty* indicates that an attribute is not associated with a concept. I searched concept net for attributes that represent the target properties and use the relations to retrieve positive as well as negative examples. The stereotype dataset collected by Veale (2013) also allows for such a direct search. It contains concepts that serve as particularly good property-examples derived from corpus data. This direct search resulted in relatively small sets of examples with uncontrolled distributions.

Taxonomic Categories

To increase the number of candidates in a way that increases the chances of diverse examples within the positive and negative class and high similarity of positive to negative examples, I exploit taxonomic category information captured in the Princeton WordNet noun hierarchy Miller (1995). Consider the example of the property **fly**: There are several semantic categories whose members are likely to have the property fly, such as BIRD, VEHICLE, and INSECT. These categories do, however, also contain negative examples of the property, such as flightless birds and insects, and vehicles that do not fly. The advantage of such negative examples is that they are likely to share a many other properties with their positive counter parts (e.g.

used_for_transportation and **wings**). Beyond increasing the number of candidates, the advantage of picking more than just a single category is an increase in diversity of candidates.

To access members of semantic categories via the Princeton WordNet hierarchy, I exploit its hyponymy relations. For each property, I manually select words that express suitable semantic categories. I then manually select the synsets that best represent each category (based on synset members, definition, and hyponyms) and extract all lemmas included in their hyponym synsets.

Logical Implications

In a pilot study (Sommerauer and Fokkens, 2018), we manually verified negative property examples extracted from the CSLB norms. Rather than counting every concept for which a property has or has not been listed as a positive or negative example, we exploited logical implications between properties to preselect concepts that were highly likely positive or negative examples of a target property. In a second step, we manually verified the preselected examples. For example, we used category membership to extend the positive examples of the property **is_a_bird** by means of selecting all concepts for which the property **is_an_animal** has been listed. Vice-versa, we assumed that concepts labeled with **has_wheels** as highly unlikely to be positive examples of **used_in_cooking**.[2] I used the resulting annotations to extend the sets of positive and negative examples of the property datasets.

It should be noted that not all properties from the initial pilot study are part of the current dataset.[3] The candidate examples collected in this process have the disadvantage of a lack of diversity. Furthermore, positive and negative examples tend to be taken from radically different semantic categories. They do, however, have the advantage of constituting reliable positive or negative examples.

4.4.2 Extraction from a Distributional Model

The strategies outlined in the previous section returned candidates for positive and negative property-examples to varying degrees of success. For instance, I collected 105 probably positive and 256 probably negative example candidates for **black**, but only 6 probably positive and 63 probably negative candidates for **round**. In addition to resulting in limited candidate sets, not all strategies guarantee a suitable example distribution. In this section, I describe how I use a distributional model to (1) increase the number of candidates and (2) specifically target negative examples with a high semantic similarity to positive examples.

Particularly challenging examples for diagnostic experiments are positive examples that cannot easily be distinguished from negative examples based on low similarity (or high distance) between them. For instance, the word *penguin* is likely to have a high distributional similarity to several positive examples of the property **fly**. If *penguin* can successfully be distinguished from positive examples (e.g. *puffin*, *seagull*, *pigeon*) in a diagnostic experiment, this is good evidence that the positive examples carry property-specific information.

[2] I carried out the selection and manual verification process together with my co-author Antske Fokkens.

[3] The pilot dataset and a record of the selected implications and the annotation discussion can be found at `https://cltl.github.io/semantic_space_navigation`.

To target particularly challenging examples, I employ the following strategy: I use verified positive examples of a property as seed words for a vector representation of the property. The verified positive examples are examples for which a property has been listed in a feature norm set or that have been verified manually. The property vector is created by taking the centroid vector of the seed word representations in a distributional model. The words close to this vector have a high cosine similarity to the positive seed words. Some of the words close to the vector may share the target property, while others will not. I extract the 200 nearest neighbors of the property vector. To specifically target challenging examples, I apply the following filtering step: I manually inspect all words that have a higher distance to the property vector than the word with the highest distance that is a verified positive example. From these words, I exclude all negative examples of the property, as they are unlikely to constitute challenging examples. For this candidate extraction step, I used a skip-gram embedding model trained on the full Wikipedia dump from August 2018 using the settings recommended by Levy et al. (2015).

4.4.3 Sampling Candidates

The candidate extraction steps presented in Section 4.4.1 and Section 4.4.2 result in large sets of candidate concepts for most properties. Most of the sets are larger than necessary and annotating the full set of candidates would result in high annotation costs. In addition, the candidate set extension via the distributional model partly returned noisy data (e.g. unconventional spelling variants). In this section, I outline how I reduced the sizes of the candidate sets while attempting to keep concept candidates representative of various linguistic factors that may impact (1) the distributional representations of words and (2) the behavior of annotators.

Preprocessing

To avoid noise, I filter the candidate concepts with respect to the following criteria:

- Is the candidate concept a noun in Princeton WordNet?
- Does the Spacy lemmatizer (Honnibal and Montani, 2017) recognize the candidate concept as a noun?

Only words for which both criteria hold are considered for further sampling.

Sampling

There are various linguistic and distributional factors that can impact the nature of distributional representations as well as the behavior of annotators (e.g. word frequency and ambiguity). To ensure that the concept candidates of a property-dataset do not over-represent one of these aspects and thus distort the results of the annotation process or the diagnostic experiments, I subsample candidates with respect to factors with a potential impact. Ideally, the positive and negative examples of each property dataset should be balanced with respect

to these factors. I first present the different factors and then describe how I use sampling to achieve a balanced distribution.

Distance to the property vector. As outlined in Section 4.4.2, I use the verified positive examples to create an approximated representation of the property in the distributional space. Words close to this vector have a high cosine similarity to the positive examples. It is likely that they are positive examples of the property or challenging negative examples. There may, however, also be positive examples farther away from the property-vector. These examples are also valuable, as they are likely to be dissimilar to most other positive examples and thus constitute particularly challenging positive examples. To ensure that the entire spectrum is covered, I sample candidates with respect to their cosine distance to the property vector.

Frequency. Word frequency in a corpus is a major component of the distributional characteristics of a word. It determines in how many contexts a word can appear and thus how much distributional evidence a corpus contains about its meaning. At the same time, word frequency has also been shown to have an impact on how humans process words (Brysbaert et al., 2018) and may thus have an impact on how annotators annotate concept-property relations. The positive and negative examples of each property should thus cover a broad range of word frequencies. I sample from three different frequency ranges. The ranges are determined by using a logarithmic scale.

Lexical ambiguity. If a word form is associated with more than a single meaning, it is ambiguous (e.g. *bank*). While this definition can also include homophones (i.e. words that have the same phonological realization), I limit it to equivalent spelling, as the data underlying the distributional models consist of written texts. Lexical ambiguity can describe ambiguity in terms of word senses (i.e. one form, multiple senses) and ambiguity in terms of reference (i.e. pronouns can be used to refer to different person-entities). I only consider ambiguity in terms of senses, as it is recorded in lexical resources. Ambiguity has a major impact on context-free distributional models, as words with different senses will be used in different (and potentially unrelated) contexts. It has been shown that ambiguous words receive vectors that place them in between the different usages in a semantic space (Del Tredici and Bel, 2015). Contextualized language models are equipped to distinguish different usages of the same word form. Initial experiments on contextualized models indicate that rather than representing neatly distinguishable senses of polysemous words, contextualized models are also impacted by ambiguity and form semantic regions that are not clearly distinguishable in terms of senses (Yenicelik et al., 2020). In addition, ambiguous lexical items are likely to impact the behavior of annotators in annotation tasks. This is relevant for the collection of concept candidates, as the diagnostic dataset requires fine-grained semantic annotations of the relations that hold between properties and concepts (see Chapter 3).

Ambiguity can be caused by different phenomena: A major cause of lexical ambiguity is polysemy. Polysemous words have more than a single sense. Lexical resources usually mark this by means of multiple definitions of a lemma. In Princeton WordNet, the same lemma can be part of multiple synsets. Some resources distinguish polysemy from homonymy (e.g. the Longman Dictionary of Contemporary English (LDOCE) (Proctor, 1978)). The senses of polysemous words are considered to have emerged from the same sense and are thus semantically related (e.g. *tree* in the sense of a plant and *tree* used to describe a diagram). In

contrast, homonyms are considered to be accidental and historically as well as semantically unrelated (e.g. *bank* in the sense of the institution compared to *bank* in the sense of river bank). Given this difference, it can be expected that homonyms may link radically different word senses with the same form. Thus, the impact of homonymy on a context-free distributional space may be more extreme than that of polysemous words.

Polysemy can be divided into different types that may impact the context free models differently: Polysemy can be caused by metaphorical mappings between conceptual domains and thus affect entire groups of words (Lakoff and Johnson, 1980). For instance, abstract phenomena, such as emotions, are often described in terms of physical phenomena (*explode of anger*, *boiling blood*) (e.g. Kövecses, 2000). Such mappings are traditionally not recorded in lexical resources, but have been studied by means of corpus annotation. I extract words with metaphorical uses from the MIPVU corpus (Steen, 2010). Next to metaphorical mappings, polysemy can be caused by metonymy. In contrast to metaphor, metonymy tends to connect senses that are related more closely. For instance, the noun *chicken* can be used in the sense of the bird, but also in the sense of a type of meat. Compared to metaphorical senses, metonymic senses can be expected to share a higher number of usage patterns.

To distinguish the different phenomena involved in lexical ambiguity listed here, I apply the following strategy: I use the difference in entries in the LDOCE dictionary to distinguish homoymy from polysemy. To detect polysemous words involved in metaphorical mappings, I exact metaphorically used nouns from the MIPVU corpus. Words marked as polysemous in the LDOCE dictionary that are not annotated as metaphorical expressions in the MIPVU corpus are counted as polysemous words that are likely to be caused by metonymy or other phenomena. Words with only a single sense in the LDOCE are considered monosemous. This strategy does not guarantee a completely accurate classification of polysemous words, but it constitutes an approximation. Sampling from the different types of lexical ambiguity described here increases the chances of a balanced representation of lexical ambiguity. I present a validation of this approximated classification in Section 4.5.

Psycho-linguistic factors. Several lexical phenomena can potentially impact the behavior of human annotators when confronted with individual lexical items. For example, people can be expected to react differently to expressions they are highly familiar with than to expressions that they hardly know. Similarly, words with concrete interpretations may have a different effect from words with abstract interpretations, in particular when considering semantic properties. The MRC psycholinguistic databased (Coltheart, 1981) captures such factors by means of human ratings. It also includes information about the average age at which a word tends to be acquired. While this by itself may not necessarily impact annotator behavior, it is likely to be correlated with familiarity and possibly concreteness. For each of the three aspects, I sample from different ranges of the ratings (or age information) to achieve a balanced distribution.

All factors considered in the sampling process are summarized in Table 4.3. Each factor either compasses a spectrum of numerical values (e.g. frequencies) or categories. To sample from different ranges, I create three equally distributed bins over the entire vocabulary of concepts that can then be used for sampling. In the case of categories, I simply sample from the categories. To fill the positive and negative (or at this point undecided) classes of each

property dataset, I iteratively draw concepts from the different bins associated with each factor.

factor	source	bins	impact repr.	annot.
distance to property vector	Wikipedia skipgram model	3 based on distance	✓	✓
frequency	Wikipedia corpus	3 based on log frequency	✓	✓
ambiguity	LDOCE dictionary, MIPVU corpus	homonymy, metaphor, polysemy (other), monosemy	✓	✓
concreteness	MRC database	3 (base on score)		✓
familiarity	MRC database	3 (based on score)		✓
age of acquisition	MRC database	3 (based on score)		✓

Table 4.3: Factors used for sampling.

4.5 Overview and Validation

In this section, I present an overview of the resulting dataset and a validation of the method I used to categorize words according to different degrees of lexical ambiguity. Table 4.4 presents an overview of the resulting property datasets. The table shows the number of candidate concepts per class. For several properties (e.g. **roll** and **sweet**), the number of likely positive or likely negative cases was low. In such cases, most candidates were sampled from the undecided class.

To validate the categorization of words into different types of lexical ambiguity, I test the following assumtions:

- Semantic similarity metrics should reflect the different degrees of ambiguity. Senses of homosemous words should have the lowest similarity; senses related by metonymy should have the highest similarities.
- Homonyms and words with metaphorical senses should have the highest changes of having both a concrete and an abstract sense. This shift should be less common among metonymic senses.

I test these assumptions by means of the Princeton WordNet noun hierarchy and cosine similaritty in the Wikipedia skip-gram model. To measure semantic similarity between synsets, I use (a) the graph based-similarity measure developed by Wu and Palmer (1994) (called wup) for the distance between synsets and (b) the average cosine distance of synset

property	pos	neg	pos/neg	total
warm	20	28	118	166
hot	19	20	108	147
red	46	59	69	174
square	6	23	90	119
green	57	58	60	175
cold	18	22	81	121
sweet	28	1	145	174
blue	22	60	61	143
yellow	45	65	64	174
round	37	2	101	140
black	60	58	34	152
juicy	20	6	148	174
swim	57	61	62	180
roll	4	1	115	120
lay_eggs	61	61	32	154
fly	58	61	61	180
dangerous	63	61	17	141
used_in_cooking	59	60	60	179
wheels	54	16	45	115
wings	58	60	29	147
made_of_wood	59	12	81	152

Table 4.4: Overview of dataset size after sampling.

members in the distributional model. To approximate the cosine similarity of synsets, I extract the monosemous lemmas from each synset and calculate their mean cosine distance in the distributional model. To test the assumption about concrete and abstract senses, I check whether the synsets associated with a word are subsumed under the abstract and the concrete part of the WordNet noun hierarchy. In addition to these metrics, I also present the mean number of WordNet synsets per word.

The results of the validation are presented in Table 4.5. All similarity measures confirm the assumptions about semantic similarity between senses. The analysis of abstract and concrete senses is also in line with the expectation. Furthermore, it can be observed that homonyms and metaphorical words tend to have more WordNet synsets than words affected by metonymy. The table also shows the results for monosemous words as a sanity check: As expected, they tend to have the lowest number of WordNet synsets. If there is more than a single synset, the semantic similarity between the synsets is considerably higher than for all other ambiguity types.

4.6 Summary

In this chapter, I have presented core properties of the methodology used to diagnose semantic properties in distributional models. These properties impose specific requirements on a diagnostic dataset. A good diagnostic set for a particular semantic property should consist of

type	**n syns**	**wup sim**	**min wup sim**	**cos syns**	**abstract-concrete shift** (%)
homonymy	8.28 (6.97)	0.32 (0.16)	0.20 (0.18)	0.20 (0.20)	60
metaphor	8.32 (7.68)	0.35 (0.19)	0.24 (0.21)	0.24 (0.23)	45
metonymy (approx.)	3.01 (2.72)	0.53 (0.32)	0.48 (0.35)	0.57 (0.38)	24
monosemy	1.97 (2.43)	0.78 (0.32)	0.76 (0.35)	0.80 (0.31)	09

Table 4.5: Validation of ambiguity types by means of semantic similarity measures and synset distribution in the Princeton WordNet hierarchy. Averages have been calculated on nouns only (standard deviation in parentheses).

a diverse set of positive examples and have negative examples that are overall similar to the positive examples. Distinguishing positive from negative examples should require identifying the target property.

The methodological considerations inform the selection of properties and concepts for the diagnostic dataset. I have used various lexical resources in combination with a context free distributional model to collect candidates for positive and negative property-examples. The candidates still need to be verified and annotated with property-concept relations (Part III).

Both the representation of concepts in a distributional model and the annotation process may be impacted by various distributional and linguistic factors (word frequency, similarity to confirmed positive examples, ambiguity, psycholinguistic factors). I consider these factors when selecting the candidate concepts for annotation. Different degrees of ambiguity are approximated by means of combining information from different linguistic resources. I have presented a validation of this strategy by means of testing whether different semantic distance measures can reflect the different degrees of ambiguity. The results are in line with the assumptions, which indicates that the selected candidates should indeed cover a broad range of lexical ambiguity.

Part III

Dataset

Part IV of the thesis focuses on the compilation of a diagnostic dataset following the model proposed in Chapter 3 and the architecture proposed in Chapter 4 by means of crowd annotation. The part is divided into three chapters: Chapter 5 introduces the annotation task used to collect semantic judgments from crowd annotators. In addition to the task design, it provides information about quality checks and worker managements, as well as different annotation cycles. Chapter 6 presents an approach towards evaluating crowd annotations for a semantic task that is expected to trigger disagreement. Rather than using disagreement as the sole indicator of quality, the chapter proposes an approach which tests whether disagreement follows expected patterns and uses a task-inherent, agreement-independent measure to establish annotation quality. The final chapter (Chapter 7) analyses to what degree the collected dataset of properties, concepts, and relations is suitable for diagnostic experiments.

5. Annotation Task

5.1 Introduction

This chapter presents the design and procedure of the annotation task used to compile a diagnostic dataset of properties, concepts, and their relations. The goal of the annotation task is to annotate properties and candidate concepts with fine-grained semantic relations. For instance, the dataset should contain information about the fact that lemons tend to be yellow and that the property **yellow** is typical of lemons (e.g. **yellow**-*lemon*-`typical_of_concept`). The relations reflect different factors that may impact whether property-information tends to be made explicit in corpus data based on theoretical and empirical research (Chapter 3).

Annotating properties and concepts with fine-grained semantic judgments encompasses several challenges: First and foremost, the task should cover a selection of properties and a substantial number of concepts per property to ensure that the resulting datasets allows for diagnostic experiments. Beyond this, a task design for a semantic annotation task should anticipate semantic phenomena such as vagueness and ambiguity that may trigger disagreement. To address these two challenges, I opt for crowd rather than expert annotation. In a crowd annotation set-up, the task can be distributed over many annotators. Consequently, it is possible to collect a large volume of data. In addition, it is possible to collect multiple different judgements per annotation unit. The distribution of judgements for individual annotation units could reflect different semantic phenomena in the data, such as ambiguity. While lexical ambiguity is not the focus of this thesis, it is part of lexical data and likely to impact distributional representations (Del Tredici and Bel, 2015, e.g.). Annotation units containing ambiguity should trigger disagreement between annotators, while clear-cut units should lead to agreement.

Crowd annotation entails that semantic judgments will be collected from non-experts. This means that the task should be designed in such a way that it requires no training (apart from simple instructions). Furthermore, the task should ideally consist of small annotation units that can be judged quickly and based on intuition. Another consequence of crowd annotation is having limited control over annotators; while crowd platforms allow for some degree of selection, it is difficult to avoid collecting low-quality annotations. At the same time, some crowd workers deliver high-quality annotations and become increasingly skilled at the task over time and should be motivated to keep working on the task.

To address the specific challenges of using crowd annotation for a rather complex semantic task, I take the following steps: To make the task accessible for untrained annotators, I translate individual semantic relations to statements that describe a specific semantic relation between a property and a concept (outlined in Section 5.2). To keep the cognitive load low, a minimal annotation unit is defined as a property-concept-relation combination expressed as a single statement. For instance, annotators are shown the following statement (expressing the

combination of **red**, *blood*, and `typical_of_property`) and asked to indicate whether they agree or disagree with it:

(12) "Blood" is one of the first things which come to mind when I hear "red" because (a/an) blood is a typical example of things which are red.

The details behind the annotation process as well as strategies used to control the quality are outlined in Section 5.3. The annotation was carried out in multiple rounds and included modifications and adaptations of the task. This resulted in a number of different dataset versions (outlined in Section 5.4). The full dataset, including information about the task set-up can be downloaded from the following Github repository: `https://github.com/PiaSommerauer/PropertyConceptRelations`. The repository allows to trace changes in the task presentation between all annotation cycles.

5.2 Statement Generation

A core component of the annotation task is the translation of property-concept relations to natural language sentences that can be judged by untrained crowd annotators, ideally based on their intuition. Property-concept relations represent highly abstract ideas, such as for instance the idea that property information is a highly implied aspect of conceptual knowledge and shared across a larger semantic category. To make such semantic notions accessible for crowd annotation, I translate each relation to a natural language statement about a specific property and a specific concept. In this section, I present the statement templates for each semantic relation. The templates presented in this section are a result of adaptations between multiple annotation iterations (see Section 5.3). The publicly available dataset contains all versions of the statements.

5.2.1 Impliedness

A fundamental notion of the hypothesis framework presented in Chapter 3 is that highly implied information is usually not mentioned explicitly. This idea is represented by the property-concept relation `implied_category` and translated to the following statement:

Relation: `implied_category`

Template: I know that (a/an) [concept] [property] as most or all other things similar to (a/an) [concept] [property].

Positive example: I know that (a/an) *dragonfly* has **wings** as most or all other things similar to (a/an) *dragonfly* have **wings**.

Negative example: I know that (a/an) *wasp* is **green** as most or all other things similar to (a/an) *wasp* are **green**.

The template shown above is used to generate specific statements using properties and concepts from the collected candidates. For each property type (e.g. part properties, perceptual

properties, activities) the template is modified slightly to ensure that the statement sounds relatively fluent.[1] Examples illustrate an instance in which most crowd workers would be expected to agree with the statement (positive example) and an instance in which they would be expected to disagree (negative example).

5.2.2 Variability and Specification

A major factor hypothesized to lead to explicit property mentions is variability. If there is variation for a certain type of property (e.g. color), it can be expected to be specified explicitly. I distinguish between a limited (e.g. **red/yellow/green** *apple*, `variability_limited`), and a wide, open-ended range of possible properties (`variability_open`).

Relation: `variability_limited`

Template: You can find (a/an) [concept] which [property]. [Property] is one of a few possible [property-category] (a/an) [concept] usually has. There is only a limited range of possible [property-category].

Positive example: You can find (a/an) meat which is juicy. Juicy is one of a few possible qualities (a/an) meat usually has. There is only a limited range of possible qualities.

Negative example: You can find (a/an) cheese which is green. Green is one of a few possible colors (a/an) cheese usually has. There is only a limited range of possible colors.

Relation: `variability_open`

Template: You can find (a/an) [concept] which [property]. [Property] is one of many possible [property-category] (a/an) [concept] usually has. The range of [property-category] is almost unlimited.

Positive example: You can find (a/an) insect which is black. Black is one of many possible colors (a/an) insect usually has. The range of colors is almost unlimited.

Negative example: You can find (a/an) cabbage which is green. Green is one of many possible colors (a/an) cabbage usually has. The range of colors is almost unlimited.

5.2.3 Illustration and Typicality

Some concepts are so closely associated with a property that they can be used as illustrations of the property. This type of close relationship between property and concept is represented by the relation `typical_of_property`. A similarly close, but not equivalent relationship between property and concept can hold for properties that are very closely connected to the concept and immediately come to mind when thinking of the concept. This relationship is represented by the relation `typical_of_concept`. For concepts that serve as illustration of the property, explicit property mentions in text are expected to be more likely than for property-concept pairs in which the property merely has a strong association with the concept.

[1] In this section, I limit the discussion of statements to default templates. Variations used for part properties and scalar properties are provided in Appendix .

Relation: `typical_of_property`

Template: "[Concept]" is one of the first things which come to mind when I hear "[property]' because (a/an) [concept] is a typical example of things which are [property]'.

Positive example: "Sugar" is one of the first things which come to mind when I hear "sweet' because (a/an) sugar is a typical example of things which are sweet'.

Negative example: "Orchid" is one of the first things which come to mind when I hear "blue' because (a/an) orchid is a typical example of things which are blue'.

Relation: `typical_of_concept`

Template: "[Property]" is one for the first things which come to mind when I hear "[concept]' because [property] is one of the typical [property-category] of (a/an) [concept]'.

Positive example: "Juicy" is one for the first things which come to mind when I hear "peach' because juicy is one of the typical qualities of (a/an) peach'.

Negative example: "Hot" is one for the first things which come to mind when I hear "flowerpot' because hot is one of the typical temperatures of (a/an) flowerpot'.

5.2.4 Afforded Actions

Certain properties are particularly important for concepts because they enable actions or uses. This property-concept relationship is represented by the relation `affording _activity`.

Relation: `affording_activity`

Template: I know that [property] is necessary for many things (a/an) [concept] does or is used for.

Positive example: I know that having (a/an) wings is necessary for many things (a/an) gull does or is used for.

Negative example: I know that being blue is necessary for many things (a/an) buzzard does or is used for.

A number of properties express activities themselves directly (e.g. **swim**, **roll**). Activities that instances of a concept usually engage in are expected to arise from distributional data (`afforded_usual`). In contrast, activities that instances of a concept are able do perform but do not usually engage in are not expected to be mentioned systematically enough to arise from co-occurrence patterns (`afforded_unusual`).

Relation: `afforded_usual`

Template: I know that all or most [concept] [property] regularly or are used for [property] regularly.

Positive example: I know that all or most crow(s) fly regularly or are used for flying regularly.

Negative example: I know that all or most lorry(s) fly regularly or are used for flying regularly.

Relation: `afforded_unusual`

Template: All or most [concept] can [property]/be used for [property]. This is not what they normally do or are used for.

Positive example: All or most bulldog(s) can swim/be used for swimming. This is not what they normally do or are used for.

Negative example: All or most ship(s) can swim/be used for swimming. This is not what they normally do or are used for.

5.2.5 Negative relations

To facilitate the annotation process, I include a range of negative relations. Annotators are likely to have different personal thresholds for indicating that a property cannot apply to a concept. The range of negative relations includes rare property-concept combinations (`rare`), unusual property-concept combinations (`unusual`) and impossible combinations `impossible`).

Relation: `rare`

Template: I think (a/an) [concept] can be [property], but this is rare or uncommon.

Positive example: I think (a/an) cheese can be sweet, but this is rare or uncommon.

Negative example: I think (a/an) notebook can be square, but this is rare or uncommon.

Relation: `unusual`

Template: Usually, (a/an) [concept] is not [property], but there could be a highly unusual situation in which (a/an) [concept] is [property].

Positive example: Usually, (a/an) corncob is not red, but there could be a highly unusual situation in which (a/an) corncob is red.

Negative example: Usually, (a/an) milk is not cold, but there could be a highly unusual situation in which (a/an) milk is cold.

Relation: `impossible`

Template: I think it is impossible for (a/an) [concept] to [property]/be used for [property].

Positive example: I think it is impossible for (a/an) pig to lay_eggs/be used for laying eggs.

Negative example: I think it is impossible for (a/an) tea to be sweet.

5.2.6 Creative and Metaphorical Properties

Finally, the task includes the possibility to mark metaphorical or creative language use (`creative`). The intention behind this relation is to detect clear cases of non-literal combinations. While such combinations may constitute interesting data for research on non-literal language use, they pose the risk of introducing noise in a diagnostic experiment. The example involving the pair **green**-*newspaper* shown below illustrates such an instance: In this statement *green* can indicate a political orientation rather than a visible color; while newspapers can have political orientations, they are unlikely to have a visible, green color.

Relation: `creative`

Template: I could say (a/an) [concept] is [property], but I would most certainly not mean it literally.

Positive example: I could say (a/an) newspaper is green, but I would most certainly not mean it literally.

Negative example: I could say (a/an) taxicab has wheels, but I would most certainly not mean it literally.

5.3 Annotation Task and Process

This section outlines the details of the annotation task design, annotation process and participant management.

Task At its core, the annotation task is framed as a binary decision task: Given a statement, participants are asked to decide whether they agree or disagree. Participants are shown one statement at the time, accompanied by two examples statements; one illustrating a statement they are most likely to agree with and one illustrating a statement they are most likely to disagree with. An example can be seen in Figure 5.1. The specific examples shown with each annotation unit can be found in the dataset repository. Participants were instructed to read statements carefully and look up words they did not know. A single batch of annotations completed by a participant consisted of around seventy statements and was estimated to take about seven minutes in total (each sentence should take between seven and ten seconds).

Annotation platform and participant recruitment The annotation task was set-up using Lingoturk; a tool designed for setting up different types of (psycho-)linguistic tasks developed by Pusse et al. (2016). The tool offers easy connection to commonly used participant recruitment platforms, such as Amazon Mechanical Turk or Prolific. For this task, Prolific was chosen.[2] Peer et al. (2017) show that the annotation quality of annotators recruited via Prolific is higher than for Amazon Mechanical Turk workers. The platform encourages fair payment and asks researchers to pay participants based on the time they estimate for a task rather than per annotated item.

[2] https://www.prolific.co/

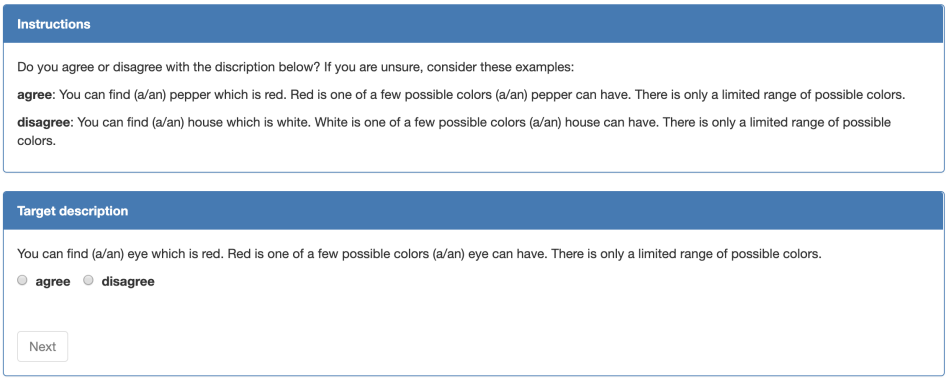


Figure 5.1: Example of an annotation unit shown to annotators in the task interface.

Initially, the task was open to all participants who indicated fluency in English. At a later stage, I decided to limit the task to annotators who had previously delivered high-quality work using Prolific's 'allow-list' option. Participants were paid based on UK minimum wage per hour regulations. For each batch of annotation units, the duration and reward were estimated based on the number of statements in the batch and the duration of previous annotation batches. Each annotation batch was annotated by ten workers.

Quality control To control whether annotators understood the task and delivered serious responses, I used two quality control strategies: check questions and task-inherent consistency checks. This section explains how these strategies were used.

Each batch contained two statements containing clear and unambiguous property-concept combinations (see Example 13). Competent speakers of English who read the instructions should all give the same answer to these questions. While such checks can provide first indications, they also encompass disadvantages: Checks have to be updated continuously, as the same annotator can work on several annotation batches and may notice repeated questions. In the worst-case scenario, attention check questions are identified by participants and shared among crowd workers to avoid being identified as untrustworthy. Furthermore, attention checks only provide information about two out of around seventy judgments.

(13) "Pink" is one for the first things which come to mind when I hear "grass' because pink is one of the typical colors of (a/an) grass'.

To address the limitations of check questions, I resorted to using task-inherent consistency checks. The underlying assumption behind these checks was that annotators should not give contradictory judgments. If they indicate that a property applies to most or all instances of a concept, they should be consistent in their judgments. Thus, they should not click on 'agree' for any of the statements expressing a negative relation between the same property-concept pair. As the task may contain ambiguous or difficult instances, a single contradiction is not necessarily a sign of bad quality. Multiple contradictions, however, can be seen as an

indication of low-quality work. This is particularly obvious if property-concept pairs did not trigger contradictions in the majority of annotators.

The two control mechanisms were used as follows: In a first step, I inspected answers of participants who failed both attention checks or delivered a higher number of contradictory responses than most other annotators on the same batch. In cases in which multiple responses could not be justified (e.g. based on ambiguous instances in the annotation units), I rejected the submission via Prolific and did not pay the participant. The task was then reassigned to another annotator. This only occurred in rare cases.

In a second step, I recorded attention check fail rates and contradictions of annotators over all batches that annotators had worked on. Once I had recruited a group of participants who had worked on multiple batches and delivered high-quality work, I added them to an 'allow-list'. Participants with high fail- and contradiction rates were removed from the list. Subsequent annotation batches were only made available to participants on the allow-list.

5.4 Dataset Versions

The final dataset is the result of multiple iterations of annotations. Each iteration was followed by an assessment of the task and data and led to adaptations. In this section, I present an overview of the annotations. Overall, the dataset is a result of six annotation cycles (summarized in Table 5.1).

iter.	task	dataset	properties
1	binary/scale	pilot blackbox	properties with at least 20 concepts in the CSLB norms
2	statement judgment	discarded	red, round, roll
3	statement judgment	diagnostic dataset	red, round, roll
4	statement judgment	diagnostic dataset	wheels, made_of_wood, hot, square, dangerous, lay_eggs, yellow, fly, sweet, black
5	statement judgment	diagnostic dataset	wings, warm, used_in_cooking, swim, juicy, green
6	statement judgment	diagnostic dataset	warm, hot, cold

Table 5.1: Overview of annotation cycles.

Pilot blackbox The first version of the dataset was compiled for a pilot experiment on word embedding representations (see Chapter 8).[3] This pilot dataset consists of properties and concepts derived from the CSLB feature norms (Devereux et al., 2014).

We select features from the CSLB norms that are associated with at least 20 concepts. In an exploratory experiment, we count all concepts for which the target feature is listed as

[3]The study and dataset collection were conducted in collaboration with Antske Fokkens. The text in this section is taken from our jointly written paper (Sommerauer and Fokkens, 2018).

positive examples and all other concepts as negative examples. However, the fact that people did not list a property does not necessarily mean that a given concept is a negative example of it. For instance: *falcon* is described by *is_a_bird*, but not by *is_an_animal*.

For proper evaluation, the CSLB dataset should be extended with verified negative examples. We apply two methods to add both positive and (verified) negative properties to CSLB. First, we select properties that necessarily imply the target property (e.g. *is_a_bird* implies *is_an_animal*) or necessarily exclude the target property (e.g. *is_food* almost certainly excludes *has_wheels*). We both manually inspect the extended sets of positive and negative examples per selected property to exclude remaining noise independently, resolving disagreements after discussion.[4]

The resulting dataset has the disadvantage that negative examples largely consist of the same specific categories, e.g. negative examples of *has_wheels* are food, animals and plants. Based on these examples, we cannot tell whether the classifier performs well because embeddings encode the property of having wheels or because it can distinguish vehicles from food, animals and plants. We therefore need to expand the dataset so that it includes diverse negative and positive examples and preferably positive and negative examples that are closely related in semantic space.

Ultimately, we want to verify and increase the entire dataset and distinguish between things that always or typically have a property (e.g. **has_wheels**-*bike*, **is_yellow**-*banana*), things that can have a property (e.g. **is_pink**-*bikini*, **made_of_metal**-*plate*) and things that do normally not have a property (e.g. **does_kill**-*grape*, **is_pink** *beer*). We set up a crowdsourcing task in which we ask participants whether a property applies to a word. Possible answers are *yes*, *mostly*, *possibly* and *no*.

This crowdsourcing method has currently been applied to a selection of property-concept pairs that were labeled as false-positives by at least one of our approaches in the initial setup. In addition, we extend the property-concept pairs given to crowd workers by collecting the nearest neighbors of the centroid (calculated over positive examples of a property) and a number of seed words. We aim to (1) identify negative examples that have a high cosine similarity to positive examples in the dataset and (2) include a broader variety of words. This nearest-neighbors strategy explicitly aims at collecting words that are highly similar to positive examples of a property but are not associated with it. For instance, in order to extend the concept set for the property *has_wheels*, we used the seed words *car*, *sledge*, and *ship*.[5]

In the experiments reported in this study, we only consider properties that clearly apply to a concept as positive examples (the *yes* and *mostly* cases) and properties that clearly do not apply as negative examples, leaving disputable cases and the cases that *possibly* apply for future work. We manually checked cases of disagreement in the crowd data and selected or removed data based on these criteria.[6]

[4] All annotations, guiding principles as well as notes about resolving discussions can be found at `https://cltl.github.io/semantic_space_navigation`.

[5] The details about our selection and full lists of seed words are provided with our code.

[6] Some differences in judgment are clearly the result of lack of knowledge (e.g. not knowing a something is an animal). The original outcome of the crowd and final resulting test are provided on the Github repository.

Diagnostic dataset The diagnostic dataset is partially based on existing annotations from the pilot dataset as well as the quantified McRae norms dataset compiled by Herbelot and Vecchi (2016). The information about positive and negative examples of properties provided in these datasets was used to limited the ranges of possible property-concept relations in the annotation of the diagnostic dataset.

The diagnostic dataset was annotated in 5 annotation cycles (iteration 2-6). Iteration 2 only encompasses three properties and acted as a pilot. The responses collected in this iteration indicated that crowd annotators had considerable difficulties with the task. As a result, I revised the statements, instructions, and examples shown to participants. From iteration 3, participant responses indicated that the task was doable. Each iteration was followed by an update of examples and small adaptations required by irregularities (e.g. fluency issues in the template generation). In addition, the check questions were updated. From iteration 5, highly abstract or difficult terms were excluded from the concepts, as such terms posed considerable difficulties to the annotators.[7] The sole purpose of iteration 6 was to complete the entire spectrum for scalar heat-properties; each example of either one of the properties **cold**, **warm**, and **hot** was annotated for all other heat properties.

5.5 Summary

This chapter presented the annotation task used to collect the diagnostic dataset. I have described how property-concept-relation combinations were translated to statements crowd annotators without expert knowledge can judge relatively quickly. Furthermore, I have provided details on how participants were recruited via the online Prolific and introduced the strategies I used to monitor crowd annotators and control the quality of annotations. Finally, I have presented an overview of the different annotation cycles used to compile the dataset.

[7]The remaining concepts were filtered by Antske Fokkens and me.

6. Evaluating Crowd Annotations

6.1 Introduction

The crowd annotation task presented in the previous chapter (Chapter 5) constitutes a difficult semantic task that cannot be expected to lead to perfect agreement between annotators. The task encompasses concepts that are ambiguous (e.g. *crane*) or vague (e.g. color boundaries such as **yellow** versus **brown**) and may lead to different responses depending on participants' knowledge about specific aspects of conceptual knowledge (e.g. yellow tomatoes are less well-known than red tomatoes, but they exist). Furthermore, the negative property-concept relations (`rare`, `unusual`, `impossible`) have specifically been designed to accommodate variation between participants. Thus, disagreement is not only expected, but can constitute a valuable signal for various semantic phenomena.

This chapter presents an evaluation of the crowd annotation task that aims to consider disagreement as a valuable signal of semantic phenomena, rather than an indication of low quality. The chapter consists of two parts: The first part of the chapter presents an evaluation of the crowd annotations compared to expert annotations with respect to expected and unexpected disagreement (Section 6.2). This part of the chapter is based on the following publication:

> Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics

The second part of the chapter focuses on individual property-concept relations (Section 6.3). The aim of this part is to assess how well crowd annotators understood the intention of and distinction between different types of relations. The annotations used for both evaluations presented in this chapter are based on a subset of all annotations collected for the diagnostic dataset. Specifically, they are taken from iteration 3 and 4 (see Chapter 5 for details about the annotation iterations).

The results indicate that the crowd annotations follow expected behavior; disagreement occurs in instances where it is justified by the data. Agreement is thus not suitable as the sole indicator of annotation quality. Instead, the evaluation against expert annotations shows that a task-specific quality based on annotation coherence yields more reliable results. The evaluation on the level of relations indicates that the crowd annotations provide accurate results for several property-concept relations. However, some of the more fine-grained

distinctions could not be made by the annotators. This should be taken into account in further analyses.

6.2 Evaluation 1: Justified and Informative Disagreement

Would[1] you say leopards are yellow? Most likely, some people would while others would not. Both interpretations are valid, as the interpretation depends on a person's boundaries for the properties 'yellow' and 'brown'. Selecting only one judgment would disregard the vagueness of the expression, a phenomenon at the heart of lexical semantics. At the same time, most people would probably agree that wine can be red without having to think about it. A high number of semantic annotation tasks is characterized by unclear, difficult, ambiguous and vague examples (Erk et al., 2003; Kilgarriff and Rosenzweig, 2000). Annotation, in particular when distributed among a crowd, has the potential of capturing different interpretations, conceptualizations and perspectives and can thus provide highly relevant semantic information. Existing evaluation and label extraction methods, however, still heavily rely on agreement between annotators, which implies a single correct interpretation. Finished datasets rarely provide indications about difficulty and ambiguity on the level of annotated units.

The explanatory power of NLP experiments that aim to evaluate or analyze models depends on the informativeness of the data. This is particularly relevant for experiments which specifically aim to understand models better, such as the tradition of diagnostic experiments (Belinkov and Glass, 2019). Traditional error analyses could also benefit substantially from test sets which contain information about phenomena with a likely impact on model performance. Furthermore, knowing whether model-errors are similar to human disagreements can yield important insights about models. For instance, an analysis of natural language inference models shows that classifiers do not necessarily capture the same type of ambiguity and uncertainty as reflected in the annotations (Pavlick and Kwiatkowski, 2019). Error analyses often require manual annotation and tend to focus on small and not representative subsections of test sets (Wu et al., 2019). We argue that the behavior of human annotators can provide rich information which should be exploited, rather than reduced to single labels. Information about (dis)agreement is a by-product of the original annotation effort and thus comes for free. It can form the basis of an error analysis or, in the case of our data, should be used to draw informative conclusions from diagnostic experiments. Such experiments crucially depend on the quality and informativeness of the underlying data (Hupkes et al., 2018).

In this section, we present an approach to crowd-annotation for a diagnostic dataset which attempts to tackle these limitations. The dataset is meant to test which semantic properties are captured by distributional word representations. The task is designed to trigger fine-grained semantic judgements of potentially ambiguous examples. The behavior of ambiguous words in distributional semantic models is not well understood and thus particularly interesting (Sommerauer and Fokkens, 2018; Yaghoobzadeh et al., 2019; Del Tredici and Bel, 2015). We

[1]The text in this section is based on a joint publication with Antske Fokkens and Piek Vossen Sommerauer et al. (2020). The evaluation framework and experimental set-up was designed collaboratively implemented by me. The paper was written in collaboration. The results from the original paper have been updated and extended. The text has been updated accordingly. The structure and text from the original paper have been adapted to fit into the framework of this thesis.

investigate to what extent existing and new quality metrics indicate annotation accuracy on the one hand and ambiguity and difficulty of annotation units on the other hand. We evaluate our task from three perspectives: (1) comparison against an expert-annotated gold standard, (2) a task-specific coherence metric independent of agreement and (3) evaluation in terms of inter-annotator agreement metrics compared to predefined expectations about agreement and disagreement. In particular, we aim to investigate (1) how we can exploit the strengths and weaknesses of various suggested metrics to select and aggregate labels provided by the crowd, (2) to what degree disagreement among workers occurs in cases where it is expected and legitimate and (3) which metrics are suitable for detecting annotation units with legitimate and informative disagreement.[2]

Disagreement has been shown to indicate ambiguous cases when measured with the CrowdTruth framework (Aroyo and Welty, 2014; Dumitrache et al., 2018). However, we are not aware of work which compares different (dis)agreement and difficulty metrics. To the best of our knowledge, there is no study which tests how well different metrics can be used to identify ambiguous annotation units in a set of units annotated in terms of expected and legitimate disagreement. We show that the metrics we use give complementary insights and can be used to filter and aggregate labels in a way that produces high-quality annotations. Despite a relatively low inter-annotator-agreement, we show that worker behavior follows our expectations about agreement and disagreement and that high-quality labels can be extracted from the annotations, in particular for cases where we expect worker agreement.

The remainder of this section is structured as follows: We provide a short description of the quality requirements of our use-case (Section 6.2.2) and the annotation task designed for it (Section 6.2.3). We present our expert-annotated gold standard in Section 6.2.4 and different quality metrics in Section 6.2.5. The results of our experiments are described in Section 6.2.6, followed by a discussion and conclusion.

relation	example
typical_of_concept	"**Spicy**" is one for the first things which come to mind when I hear "*chili pepper*' because **spicy** is one of the typical tastes of (a/an) *chili pepper*'.
typical_of_property	"*Feather*" is one of the first things which come to mind when I hear "**light**' because (a/an) *feather* is a typical example of things which are **light**'.
affording_activity	I know that having (an/an) **blade** is necessary for many things (a/an) *razor* does or is used for.
variability_open	You can find (a/an) *t-shirt* which is **white**. White is one of many possible colors (a/an) *t-shirt* usually has. The range of colors is almost unlimited.
rare	I think (a/an) *wine glass* can is **made of plastic**, but this is rare or uncommon.
impossible	I think it is impossible for (a/an) *corpse* to be **alive**.

Table 6.1: Examples of statements expressing semantic relations.

[2]The crowd and expert annotations are available at this repository: `https://github.com/cltl/SPT_crowd_data_analysis`

Vague property	The property is vague. *Usually, (a/an) leopard is not yellow, but there could be a highly unusual situation in which (a/an) leopard is yellow.*
Ambiguous property	The property is ambiguous and not disambiguated in the context of the concept and description. *You can find (a/an) chutney which is hot. (A/an) chutney is usually either hot, a bit more or less hot or the opposite of hot.*
Ambiguous concept	The concept is ambiguous and not disambiguated in the context of the property and description. *I know that (a/an) trumpeter can fly/be used for flying as most or all other things similar to (a/an) trumpeter fly.*
Odd pair	The combination of the property and concept is strange and confusing. This is always the case, regardless of the description. *You can find (a/an) recliner which is square. Square is one of a few possible shapes (a/an) recliner usually has. There is only a limited range of possible shapes.*
Odd triple	The combination of the property, concept and description is strange and confusing. *I know that being yellow is necessary for many things (a/an) buttercup does or is used for.*
Differences in conceptualization	The description asks the participant to place the concept in a conceptual system. The answer depends on the conceptual system of the participant. *I know that (a/an) arrow can fly/be used for flying as most or all other things similar to (a/an) arrow fly.*
Specialized knowledge	Answering this correctly requires specialized knowledge. It is likely that not all workers are aware of this. *I think (a/an) carrot can be red, but this is rare or uncommon.*
Imagination	This depends on how creative and imaginative a participant is. This type of disagreement only matters for confusions between negative relations (e.g. rare, unusual, impossible). *I think there is a shovel which can roll/be used for rolling, but this is rare or uncommon.*

Table 6.2: Expected reasons for worker disagreement.

.

6.2.1 Related Work: Disagreement in Annotation Tasks

Recent annotation studies recognize that ambiguity, vagueness and varying degrees of difficulty are inherent to semantic phenomena (Dumitrache et al., 2019; Aroyo and Welty, 2015; Erk et al., 2003; Kairam and Heer, 2016; Poesio et al., 2019; Pavlick and Kwiatkowski, 2019). Pavlick and Kwiatkowski (2019) demonstrate that the fundamental task of Natural Language Inferencing contains large proportions of instances with multiple valid interpretations and argue that this phenomenon is central to the task rather than an aspect which can be disregarded. Herbelot and Vecchi (2016) show that even experts disagree on a difficult semantic annotation task and that interpretations are likely to vary due to differences in conceptualizations, which are in themselves justified and cannot simply be disregarded as 'mistakes'.

Despite the central nature of phenomena triggering disagreement in annotation tasks, we are not aware of evaluation methods that do not mainly rely on agreement. Traditionally, annotations by a few annotators who worked on the same units are evaluated in terms of

Kappa scores (usually Cohens's kappa) and tasks with varying workers annotating the same units (usually crowd tasks) in terms of Krippendorff's alpha (Artstein and Poesio, 2008). The CrowdTruth framework suggested by Aroyo and Welty (2014) and Aroyo and Welty (2015) offers a more fine-grained view by distinguishing the levels of workers, units and labels, rather than reducing the entire task to a single score. The goal is to distinguish meaningful disagreements (i.e. agreements by reliable annotators) from noise (i.e. disagreement or agreement by generally unreliable annotators). The framework provides scores for workers, annotation units (clear units receive a high score, units triggering disagreement between reliable annotators a low score), labels and associations between units and labels. The scores can be used to aggregate labels and for identifying unclear annotation units, as for instance shown by (Dumitrache et al., 2015) and (Dumitrache et al., 2019). Other approaches attempt to discover disagreeing but valid interpretations in annotations based on clustering (Kairam and Heer, 2016) and Gaussian modeling (Pavlick and Kwiatkowski, 2019). While these approaches provide valuable insights, we focus on transparent and simple methods for quality assessment which do not require a large volume of data.

6.2.2 Quality Requirements for a Diagnostic Dataset

The purpose of the annotation task is to compile a dataset of properties and concepts that allows for diagnostic experiments on word embedding representations. Annotations should be provided by a crowd, rather than experts, as we are interested people's general perception rather than expert judgments. Though we focus on a task with these specific characteristics, we believe that the general approach presented in this chapter can also yield important insights in other, perhaps more traditional annotation scenarios.

Experiments in the tradition of model analysis require informative and high-quality data, as they aim to discover general tendencies about what kind of information models can capture. The diagnostic dataset proposed in this thesis is meant to test whether a semantic property (e.g. 'flying') is encoded by embedding representations or not. This can be investigated by testing whether positive (e.g. 'seagull', 'airplane', 'bee') and negative candidate concepts (e.g. 'penguin', 'train', 'ant') can be distinguished purely based on their embedding (see Chapter 4). The examples should not only be used to test whether a specific semantic property is encoded in embeddings, but, beyond this, help to uncover underlying factors determining whether a property is reflected in a distributional representation of a concept or not. Therefore, the concept-property pairs should be annotated with semantic relations reflecting various linguistic factors (see Chapter 3). Each concept-property pair can be connected by one or more of a total of ten relations (for instance expressing types of typicality or whether there can be variation in instances of concepts). The semantic relations can be grouped with respect to the subset of concept-instances a property applies to (most to all, some, or few to no instances of a concept). This enables diagnostic experiments with positive and negative examples.

We encountered the problem of annotation evaluation given expected disagreement while compiling this dataset as it contains a high number of ambiguous instances and instances of varying degrees of difficulty, for which disagreement can be valid and meaningful. In an ideal scenario, our analysis of annotations can (1) provide an overall indication of annotation quality which does not purely rely on agreement and (2) distinguish different types of disagreement.

At the most coarse-grained level, it should distinguish justified disagreement from noise (caused by mistakes or spammers).

6.2.3 Characteristics of the Annotation Task

The goal of the annotation task is to annotate property-concept pairs with relations. To make the task simple and suitable for a crowd of untrained workers, we turned it into a binary-decision task. This means that a single annotation unit consists of a property-concept-relation triple. This results in ten annotation units per concept-property pair. As the relations have rather abstract names, we translate them to natural language statements describing a property and a concept. The following statement is an example of a description expressing the property-concept-relation triple **black**-*rhino*-`variability_limited`: *You can find (a/an) rhino which is black. Black is one of a few possible colors (a/an) rhino usually has. There is only a limited range of possible colors.* Participants are asked to indicate whether they agree with a given statement about a property and a concept. More example statements are listed in Table 6.1.[3]

To avoid triggering random answers, we encourage participants to look up words they do not know. Each statement is introduced by a short instruction sentence and an example of the same relation and property-type which would most likely trigger the response 'agree' and which would trigger 'disagree'.

We used the freely available Lingoturk software (Pusse et al., 2016) to set up an annotation environment and distributed the task via the recruitment platform Prolific.[4]

We split the dataset into batches of around 70 descriptions. A worker who is proficient in English would need about 10 minutes per batch. While some statements may be difficult to judge and therefore take more time, most are expected to be rather intuitive and easy to answer. Annotators were paid based on the UK minimum wage. Each unit was annotated by 10 workers. To enable regular quality checks, we always include the full range of descriptions associated with a property-concept pair in the same batch.

This enables us to check whether answers contradict each other. It has the disadvantage that the diversity of property-concept pairs in a batch is low.

We monitored the quality of the annotations during the annotation process and used intermediate worker evaluations to 'recruit' good annotators. Rather than rejecting low-quality submissions, we developed a 'allow-list' approach (please refer to Chapter 5 for details). Prolific enables researchers to distribute studies exclusively among a pre-selected group of workers. We test whether workers contradict themselves in their answers (explained in more detail below), for instance by judging a property as typical of a concept and at the same time stating that it is unusual of the concept. As we do not know how much legitimate disagreement could be expected in a single batch, we decide to rely on an agreement-independent metric rather than inter-annotator agreement.

[3]The entire set of input data can be found at this repository: `https://github.com/cltl/SPT_annotation`

[4]`https://www.prolific.co/`

6.2.4 A Gold Standard for Accuracy and Expected Agreement

We establish a gold standard to evaluate (1) the accuracy of annotations extracted based on different quality metrics and (2) the ability of different metrics to identify justified and potentially meaningful disagreement. The authors of the paper annotated a subset of already annotated units. The units for expert annotation were selected from units with high, medium and low agreement. Agreement was established by calculating Krippendorff's alpha on the level of concept-property pairs (each pair has up to ten units).[5] The inter-annotator agreement before discussion was 0.51 and 0.72 after discussion (averaged pairwise Cohen's kappa). We count all units in which agreement between experts could not be reached as units with expected disagreement. These units (23 in total) are excluded from the gold standard for label accuracy, as there are no incorrect answers in these cases.

We also indicated whether we expected the crowd to disagree for legitimate reasons. Examples of such disagreements are shown in Table 6.2. We identify different reasons for expected disagreement, such as vagueness in the property, ambiguity in either the concept or the property, odd property-concept combinations. We used these categories to facilitate the expert annotation process. While they served as a helpful tool for annotation and discussion, the inter-annotator agreement with respect to the disagreement categories remained low. It has to be considered that in most cases, various categories interact. When discussing annotations, we could frequently reach agreement about the subset of disagreement categories involved in an annotation unit, but disagreed about where the emphasis should be placed. In our current analysis, we simplify and distinguish the following three categories: agreement, possible disagreement and almost certain disagreement. Agreement was chosen for cases where all annotators expected agreement, possible disagreement for mixed cases and disagreement for cases where all annotators indicated they expected disagreement. We argue that taking these unions is most sensible, as multiple perspectives are necessary to discover possible reasons for disagreement. In total, we expect agreement for 49 units, possible disagreement for 48 and almost certain disagreement for 57 units. For 23 of the 57 units, a gold label could not be reached in expert discussion.

6.2.5 Quality Metrics

We experiment with three types of quality metrics: We consider traditional inter-annotator agreement, quality scores in the CrowdTruth framework and our own, task-specific coherence metric. The metrics assess different aspects of the annotated dataset, as explained below.

Traditional Inter-Annotator-Agreement

Traditionally, annotation tasks are assessed in terms of inter-annotator agreement (Artstein and Poesio, 2008). Crucially, inter-annotator agreement metrics should go beyond simple ratios and account for the possibility of agreement by chance. Widely used scores which do this are Cohen's Kappa (suitable for pair-wise assessment of annotators) and Krippendorff's alpha

[5]For some pairs, some relations were excluded based on existing annotations from Herbelot and Vecchi (2016). This resulted in a set of 154 units (containing 19 property-concept pairs and 11 different properties).

(suitable for a large number of annotators who are not consistent across the set). Both scores range between -1 and 1. Artstein and Poesio (2008) argue that Computational Linguistics tasks should require an agreement of 0.8 (while agreement above 0.67 is generally considered acceptable for some tasks). Such a strict threshold would not do justice to our task, which is characterized by expected ambiguity and disagreement. Traditionally, these metrics are used to give indications about the quality of the full set. In contrast, we use them directly to investigate whether expected disagreement indeed leads to lower alpha scores.

CrowdTruth Metrics

The CrowdTruth framework was specifically designed to account for ambiguity and different levels of difficulty in a crowd-annotation setting. Beyond accounting for variation in the data, it also considers that crowd workers may have different abilities and that labels used in the annotation process can vary with respect to clarity. Rather than using a single aggregated score, the framework proposes metrics for workers, annotation units, labels and association strength between units and labels. Each task-component (workers, units and labels) is represented by a vector. The scores are calculated in terms of cosine similarities (expressing agreement) and weighted. For example an annotation unit on which most workers disagree receives a lower weight, just like a worker who frequently disagrees with other workers. Each score can take a value between 0 and 1. Dumitrache et al. (2019) show how the individual scores can be used for label identification and the identification of ambiguous units. The unit-quality-score (uqs) measures the weighted worker agreement on a particular unit and can be used to identify unclear or difficult units. The unit-annotation score (uas) measures the weighted agreement on a particular label for a unit. This indicates which label should be selected based on the analysis. Finally, we experiment with the worker quality score (wqs) for filtering low-quality workers.[6]

Task-Specific Metric: Contradiction Ratio

We define a metric specific to our task which assesses the coherence of worker judgments independent of agreement. We assume that reliable workers should not contradict themselves in the judgments of units associated with a single property-concept pair. For example, stating that a **fly** is typical of *penguin* and that it is impossible that *penguins* **fly** would count as a contradictory annotation. The semantic relations associated with a single pair can be divided into relations expressing that a property applies to all or most concept-instances, some concept-instances or few to no concept-instances. Contradictory annotations are annotations which state that relations in the `most/all`-category and the `few-none` category are true. We calculate a contradiction rate by dividing all observed contradictions by all possible contradictions for a property-concept pair. This can be done for the annotations of an individual worker or all annotations for a pair. The contradiction rate for the worker can be seen as an indication of worker quality (the lower the better). The contradiction ratio on the level of a pair can be seen as an indication of the difficulty of a property-concept combination (the higher the more difficult).

[6] We use the scores as they are defined in the appendix of (Dumitrache, 2019).

General Quality Metric: Time

Annotation platforms such as Prolific and MTurk indicate how much time participants spent on a task. These time indications can be used to identify participants who may have provided low-quality answers. According to Prolific's guidelines for accepting and rejecting submissions[7], submissions by participants who completed the task 'exceptionally fast' (i.e. 3 standard deviations below the mean) can be rejected. We evaluate to what extent comparatively fast task completion is an indication of low quality based on the time information provided by Prolific and captured by the Lingoturk platform.

6.2.6 Evaluation Against the Gold Standard

In this section, we present the results of our analysis. Section 6.2.6 presents a general overview and statistics about the collected annotations. We show the results of our evaluation against the gold standard in terms of label accuracy, followed by our evaluation with respect to expected agreement and disagreement. In Section 6.2.7, we evaluate how well different quality metrics are able to identify units with legitimate disagreement.

data	total	iter. 3	iter. 4	iter. 4
annotations	195619	20971	41447	133201
properties	13	3	3	10
pairs	1935	425	426	1501
workers	1068	285	547	455
units	17907	4105	4094	13212
workers per unit	10.92	5.11	10.12	10.08
mean number of concepts per prop.	148.85	141.67	142.00	150.10
mean number of annotations per worker	183.16	73.58	75.77	292.75
iaa_label	0.31	0.23	0.21	0.36
iaa_collapse_neg	0.33	0.24	0.26	0.38
iaa_merged	0.37	0.22	0.24	0.43
contradiction_rate_mean	0.04	0.09	0.08	0.02

Table 6.3: Overview of the dataset in total and by annotation iteration (iteration 1 and 2 are not part of the diagnostic dataset).

.

Overview

Table 6.3 shows the overview of the current state of our dataset. The table shows statistics for three intermediate versions and the total dataset. In total, we have collected almost 200 000 annotations for almost 2000 property-concept pairs covering 13 different semantic properties with on average 150 associated concepts each. On average, each worker annotated

[7]The guidelines can be found here `https://researcher-help.prolific.co/hc/en-gb/articles/360009092394-Approvals-rejections-and-returns-` (last accessed 2022/11/12).

filter	stdev	vote	f1	p	r	alpha
clean_contradictions_batch	0.5	majority	0.88	0.92	0.86	0.20
clean_contradictions_batch	0.5	uas0.65	0.87	0.87	0.87	0.20
clean_contradictions_pair	0.5	uas0.65	0.87	0.87	0.87	0.22
clean_contradictions_total	0.5	majority	0.86	0.88	0.85	0.25
-	-	uas0.7	0.84	0.84	0.85	0.19
clean_time-below_batch	1	top	0.84	0.86	0.83	0.20
clean_contradictions_pair	0.5	majority	0.84	0.86	0.83	0.22
clean_time-below_batch	1	majority	0.84	0.90	0.82	0.20
clean_contradictions_total	0.5	top	0.84	0.85	0.83	0.25
clean_contradictions_batch	2	top	0.83	0.85	0.82	0.20
clean_contradictions_batch	1.5	top	0.83	0.85	0.82	0.21
clean_contradictions_pair	0.5	top	0.83	0.84	0.82	0.22
clean_ct_wqs_batch	1	top	0.82	0.83	0.82	0.19
clean_ct_wqs_batch	1.5	top	0.82	0.83	0.82	0.19
clean_ct_wqs_batch	2	top	0.82	0.83	0.82	0.19
clean_ct_wqs_batch	0.5	top	0.82	0.83	0.82	0.19
-	-	top	0.82	0.83	0.82	0.19
exclude_contradictory_annotations	-	majority	0.82	0.84	0.81	0.24
exclude_contradictory_annotations	-	top	0.81	0.82	0.80	0.24
-	-	majority	0.81	0.86	0.79	0.19
clean_ct_wqs_batch	2	majority	0.81	0.86	0.79	0.19
clean_ct_wqs_batch	1.5	majority	0.81	0.86	0.79	0.19
clean_ct_wqs_batch	1	majority	0.81	0.86	0.79	0.19
clean_ct_wqs_batch	0.5	majority	0.81	0.86	0.79	0.19

Table 6.4: Evaluation total gold standard.

exp.	aggregation	filtering	f. unit	n_stdv	f1	p	r	alpha
agree	majority_vote	contradictions	pair	0.5	0.91	0.94	0.90	0.28
agree	uas-0.65	-	-	-	0.91	0.92	0.90	0.23
agree	uas-0.7	-	-	-	0.90	0.91	0.90	0.23
agree	majority_vote	contradictions	total	1	0.89	0.94	0.88	0.28
disagree	majority_vote	contradictions	batch	0.5	0.89	0.91	0.88	0.16
agree	majority_vote	contradictions	batch	0.5	0.86	0.93	0.84	0.28
agree	majority_vote	no_contradictions	-	-	0.86	0.93	0.84	0.32
disagree	majority_vote	contradictions	batch	1	0.85	0.89	0.84	0.15
disagree	majority_vote	contradictions	batch	1.5	0.84	0.88	0.83	0.16
disagree	majority_vote	contradictions	total	1	0.84	0.86	0.83	0.17
disagree	majority_vote	no_contradictions	-	-	0.83	0.84	0.83	0.31
agree	majority_vote	-	-	-	0.83	0.92	0.80	0.23
disagree	majority_vote	-	-	-	0.81	0.83	0.79	0.16
disagree	uas-0.65	-	-	-	0.79	0.79	0.80	0.16

Table 6.5: Evaluation of aggregated labels against expert annotations for expected agreement and disagreement in terms of precision, recall and a weighted f1-score. IAA is indicated by Krippendorff's alpha.

.

about 183 units, which is more than two batches (of 70 questions each). The total inter-annotator agreement (measured by Krippendorff's alpha) is 0.31. If relations are merged into `most-all`, `some` and `few-none`, inter-annotator agreement rises to 0.37. If just the relations in the category `few-none` are merged, the alpha score is 0.33. We improved the formulation based on the outcome of our first runs. The first two intermediate versions have lower agreement scores than the third version as a result. The number of contradictions also declines (partly due our allow-list approach).

Label Accuracy

In this section, we present the results of the evaluation with respect to the correctness of extracted and aggregated crowd annotations compared to expert annotations. We experiment with different filtering and aggregation methods using the metrics described in Section 6.2.5.

Filtering. We filter based on worker-quality metrics (wqs, contradiction rate, and time). All scores require thresholds. We experiment with different thresholds calculated in terms of n standard deviations +/- mean calculated over the entire dataset, a batch or a single property-concept pair. For time, we only considered scores below the mean. Annotations made by workers with scores outside of the threshold are removed. We vary n between 0.5 and 2 (in steps of 0.5).

Aggregation methods. We use three different strategies for aggregation: Majority vote (a relation applies if >50% of workers select 'agree'), top vote (only the relation or, in case of a tie, the relations with the most 'agree' votes per pair) and varying unit-annotation score (uas) thresholds (between 0.5 and 1 in steps of 0.05). The top vote has the limitation that it usually only selects a single relation per pair as true, which disregards the nature of the task.

Results. Table 6.4 shows the weighted f1-scores for the full set of gold annotations. In total, the set includes 131 units with a gold label (21 positive and 110 negative). The combination of filtering and aggregation methods and their thresholds results in a high number of configurations. We only report the best result for each filtering-aggregation combination.[8] All filtering methods result in full coverage for the entire gold standard set. The results show that a majority vote on labels filtered by contradiction rate on batch level yields the highest performance. The combination of the crowd truth uas metric and filtering based on contradictions performs almost equally highly. In contrast, simple majority or top vote achieve an f1-scores of 0.82 and 0.81, respectively. The best CrowdTruth method (unit-annotation-score) achieves an f1-score of 0.84, which is equally high as removing all annotations by annotators who spent less time than one standard deviation below the mean of all annotators who worked on a batch. Using the worker-quality-score to exclude annotations in combination with a top-vote only marginally improves results compared to a simple majority or top vote on unfiltered data.

We can observe that filtering based on agreement-independent factors (time, contradictions) leads to similar performances as the uas score arising from the crowd truth framework. A possible explanation for this phenomenon is that filtering annotations based on these factors fulfills a similar function as assigning lower weights to unreliable annotators. It seems that

[8] The full set of configurations and their results is included in the Github repository `https://github.com/cltl/SPT_crowd_data_analysis`.

annotators who tend to disagree with the majority (measured by crowd-truth) also spend less time on the task and tend to provide contradictory responses.

When considering the f1-scores in comparison to the total inter-annotator agreement on the evaluation set (indicated by alpha in Table 6.4), it can be seen that high performance does not necessarily depend on high agreement.

Expected Crowd Behavior

We compare the performance and inter-annotator agreement against expected agreement and disagreement. If the annotations reflect the data accurately, clear units should achieve a higher agreement than unclear, potentially ambiguous or difficult cases. Similarly, accuracy for clear cases should be high.

Table 6.5 lists the results for units in the gold set with expected agreement and the gold set with expected disagreement. In total, there are 49 units with expected agreement and 82 with expected disagreement (we merged possible and certain disagreement). For reasons of space, we only show the top three configurations, the top-configurations on the full set and some baseline configurations (majority vote on full, unfiltered set and excluding contradictory annotations). The inter-annotator agreement confirms the expected behavior (0.23 on the full set with expected agreement and 0.16 on the full set with expected disagreement). The results indicate that the contradiction-based filtering methods achieve high performance on both the set with expected agreement and expected disagreement, with only a slight advantage on the expected agreement set. The CrowdTruth unit-annotation-score (uas) methods perform highly on the set with expected agreements and drop on the set with expected disagreements (0.91 vs 0.79). We thus conclude that the contradiction-based methods provide a robust outcome and uas (CrowdTruth) can reflect differences in difficulty between sets.

A limitation of this comparison is that the two sets differ in size and balance of labels, which should be improved in an ideal set-up. The difference in inter-annotator agreement seems to be large enough to confirm that the workers behaved as expected. The results also indicate that robust labels can be extracted from a difficult set relying on contradiction-filtering.

metric	n_sd +/-mean	accuracy (disagreement)	micro f1
uqs	0	0.68	0.50
prop	0	0.71	0.48
prop_filtered	0.5	**0.68**	**0.59**
contradictions	1	0.32	0.41
contradictions_filtered	1	0.32	0.41

Table 6.6: Accuracy of different metrics in identifying units with certain disagreement. Each metric requires a threshold, which we calculate based on mean +/- n standard deviations.

.

6.2.7 Identifying Units with Valid Disagreement

In this section, we investigate whether we can identify valid disagreement and distinguish it from noise. We evaluate how well unit-based quality metrics can distinguish units with expected disagreement from units with expected agreement. For this aspect of the evaluation, we use a stricter standard for identifying expected disagreement in the expert annotations: We only use units which each of the expert annotators indicated as triggering disagreement and units with expected agreement. This leaves us with 49 units with expected agreement (as above) and 41 units with expected (and legitimate) disagreement.

We experiment with the unit quality score (uqs), proportional agreement (prop) and the contradiction rate. The latter two can be applied to the raw and filtered dataset (we use the best performing filtering method). For each metric, we calculate a threshold by establishing the mean over all units and test performance using mean +/- n * standard deviation. The best scores for each metric are reported in Table 6.6. We report the accuracy for identifying valid disagreement in comparison to the micro f1-score. The best result is achieved by using simple, proportional agreement on the dataset where contradictory annotations were removed. The contradiction rate on its own is not suitable for identifying difficult instances.

6.2.8 Discussion

In this chapter, we have attempted to fill the gap between a heavy emphasis of inter-annotator agreement on the one hand and justified disagreement on the other hand. Semantic annotation tasks have been acknowledged to contain ambiguous, difficult, vague and possibly confusing examples which are likely to trigger disagreement. While some approaches may still see these cases as marginal, we argue that they are a vital part of many linguistic phenomena and can yield important insights. In this paper, we have illustrated an approach for a dataset used in model analysis experiments. The tradition of model analysis methods places strong emphasis on the quality and soundness of datasets and the phenomena indicated by disagreement are particularly relevant for our task. However, we argue that datasets used in other experiments should be held to similarly high standards. The explanatory power of evaluation datasets for semantic tasks in general could be improved by explicitly containing information about disagreement.

We have shown that, for our particular use-case, the agreement-based metrics should not be used as the sole indicator of quality. Our results show that a task-inherent coherence check can yield important insights and serve as a valuable basis for discarding noisy annotations. While we have only shown this for our use-case, we believe that the principle can be applied to other annotation tasks as well. For example, we could imagine that the principle of logical coherence checks can be applied to a semantic role-labeling task. Predicates with contradictory semantic roles (based on the idea of selectional preferences) can be used as an indication of either noisy annotations or ambiguous annotation units. Even tasks that are particularly drawn to high disagreement, such as tasks in the domain of sentiment annotation, could benefit from such checks. In hate speech identification, it could be considered to check if (1) the same annotator uses opposing labels for very similar instances and (2) annotators completely contradict one another on the same instances (rather than just disagreeing about

the boundaries of categories (such as 'positive' and 'neutral'). We do not intend to disregard the complex nature of such a task; other contextual factors, such as the background of the annotators, can also trigger contradictions. Taking these factors into account can yield further useful insights when interpreting (differences in) annotations. We believe that considering the interaction between these factors and logical checks can provide a valuable tool for analyzing and processing annotations.

While the approach presented here can be taken as a first step, there are still a number of limitations and remaining challenges. Most importantly, it would be highly valuable if the existing metrics could be combined in such a way that we could use them for the identification of different types of disagreements. For instance, it is relevant whether workers disagree because some have more specialized knowledge than others or because the annotation unit under consideration is indeed ambiguous. It could be considered to combine different metrics in such a way that they can distinguish between disagreement due to noise, disagreement because of differences in knowledge and disagreement due to real ambiguity. A possible way to achieve this could be to use the different metrics as features in a machine learning system. This research direction would require a larger volume of expert annotated gold data.

6.2.9 Conclusion

Despite the limitations discussed above, we draw the following conclusions: (1) Absolute thresholds for inter-annotator agreement and aggregated scores over all annotations disregard the nature of a difficult semantic task with ambiguous and vague instances. Rather, evaluations should focus on whether agreement can be found in cases where agreement can be expected. Our evaluation against expected agreement and disagreement shows that worker-behavior is in line with our expectations despite overall low inter-annotator agreement. (2) The results indicate that a simple, coherence-based task-specific worker-quality check yields accurate labels, even on datasets with low inter-annotator agreement. The advantage of this check is that it does not require high volumes of data to be accurate, but can be used with only a handful of annotated units. We expect that similar checks can also be established for other tasks. Such checks can be a cheap but high-impact approach, as they can be designed in such a way that they adhere to what is important in a particular task. In our case, good workers should understand questions and not contradict themselves. This is more important than that they agree with other workers. (3) High inter-annotator agreement is not necessarily a requirement for obtaining high-quality labels. Our evaluation shows that the highest f1-score on the expert-annotated gold standard was achieved by a filtering and aggregation method which does not result in the highest alpha score on the remaining labels. (4) While our approach to the identification of legitimate disagreements is preliminary, we observe that a simple, proportional agreement metric on a dataset filtered for contradictory answers yields the best results. This research provides the groundwork for establishing the exact status of individual annotation units and thereby establish whether the information and quality is sufficient for experiments with computational linguistic models.

6.3 Evaluation 2: Accuracy of Property-Concept Relations

In this section, I[9] analyze to what extent the crowd annotations reflect valid and informative judgments on the level of individual property-concept relations. In particular, I focus on whether the aggregated labels reflect the intention behind the relations (Section 6.3.1) and to what extent crowd annotators could make fine-grained distinction between relations (e.g. distinguish between two similar notions of typicality) (Section 6.3.2). The annotations evaluated in this section are the result of the best-performing filtering and aggregation method introduced in Section 6.2.

6.3.1 Relation Accuracy

To provide a more fine-grained picture of annotation quality on the level of individual relations, I evaluated 30 randomly selected property-concept pairs for which more than 50% of annotators indicated that the property-concept relation applies. I judged each pair in terms of whether it fits the property-concept relation it is assigned to. For some instances, the annotation units leave room for interpretation. Such cases were marked as questionable.

To illustrate the evaluation procedure, consider the relation `implied_category`. The relation expresses the idea that a property is highly implied by a concept and shared across other, similar concepts that are likely to belong to the same semantic category. For example, having wheels is part of our highly implied knowledge about the concept *coach* and generally applies to the category of wheeled vehicles. For crowd annotations, the relation is expressed by the following statement:

(14) I know that (a/an) coach has (a/an) wheels as most or all other things similar to (a/an) coach have wheels.

Table 6.7 shows the 30 randomly selected property-concept pairs labeled with the relation `implied_category` by more than 50% of crowd annotators. The pairs in the table are sorted based on the proportion of annotators who agreed with the statement expressing the property-concept-relation combination ('prop_true'). The table also shows the crowd-truth score assigned to the label ('uas_true'). The final column shows my judgement in terms of correct (✓), incorrect (✗) or questionable (?). It can be observed that pairs with a high positive response rate tend to be correct, while most incorrect pairs have a low positive response rate. The highest positive response rate for an incorrectly labeled pair (**hot**: *cookware*) is 0.63. Overall, 25 out of the 30 concepts received a correct label, while 3 received a clearly incorrect label. The incorrectly labeled examples pairs are examples in which concepts are valid positive examples of the property (apples are often green, emeralds are green, cookware can be hot), but do not fulfil the requirements posed by the relation `implied_category`.

A summary of the outcome of this evaluation for all relations is shown in Table 6.8. The individual pairs and judgements per pair can be found in Appendix . The table shows the number of correct, incorrect, and questionable pairs (out of 30) per relation. To provide

[9]This component of the evaluation was carried out by me as a complementary analysis to the results presented in Section 6.2.

prop	concept	prop_true	uas_true	acc.
swim	cobia	1.00	1.00	✓
wheels	locomotive	1.00	1.00	✓
used_in_cooking	taco	1.00	1.00	✓
made_of_wood	joist	1.00	1.00	✓
wings	pintail	1.00	1.00	✓
lay_eggs	tanager	1.00	1.00	✓
fly	jet	1.00	1.00	✓
wings	archaeopteryx	1.00	1.00	✓
swim	grindle	1.00	1.00	✓
swim	duckling	0.90	0.90	✓
swim	cichlid	0.88	0.89	✓
juicy	melon	0.88	0.89	✓
made_of_wood	plank	0.88	0.87	✓
wheels	snowplow	0.88	0.87	✓
round	pancake	0.86	0.91	?
dangerous	warhead	0.86	0.87	✓
lay_eggs	pickerel	0.83	0.83	✓
roll	hubcap	0.75	0.72	✓
round	beet	0.71	0.79	✓
wings	drone	0.71	0.72	✓
made_of_wood	deck	0.67	0.67	✓
juicy	aubergine	0.67	0.66	?
dangerous	pentobarbital	0.67	0.67	✓
swim	wolf	0.62	0.64	✓
red	heart	0.62	0.62	✓
hot	cookware	0.62	0.63	✗
green	emerald	0.56	0.55	✗
square	newspaper	0.56	0.55	✓
lay_eggs	crane	0.56	0.56	✓
green	apple	0.55	0.54	✗

Table 6.7: 30 examples of the relation `implied_category`.

insights into the degree to which the positive response rate can reflect certainty, the table also shows the highest positive response rate of an incorrectly labeled pair. In addition, the number of correct pairs out of all pairs that received a perfect positive response rate of 1 (i.e. all annotators agreed).

The results show that the relations differ with respect to accuracy; while `typical_of_concept` yielded 30 correctly labeled pairs, `typical_of_property` only yielded 13 and `afforded_unusual` only 18. For all other relations, at least 22 pairs could be identified as clearly labeled correctly. When considering the role of the response rate, it can be observed that overall, incorrectly labeled pairs tend to have a comparatively low positive response rate (e.g. 0.63 for *implied_category*). For the two relations that score lowest in terms of accuracy, however, the highest positive response rates of an incorrectly labeled pair are comparatively high (0.85 and 0.86). The pairs with perfect response rates are almost always clearly correct.

The manual evaluation of randomly selected pairs thus indicates that relations differ with

respect to how well they were judged by the crowd. Complex semantic phenomena expressed by the relations `typical_of_property` and `afforded_unusual` were judged least reliably. Overall, the evaluation per relation indicates that most incorrectly judged pairs are still correct in terms of the binary class they will be assigned to (i.e. concepts in pairs assigned to positive relations are indeed positive examples of the property, concepts in pairs assigned to negative relations are indeed negative examples of the property). Exceptions exist (e.g. **blue**-*flame* is assigned to the relation `unusual`).

relation	✓	?	✗	pos response rate of highest incorrect pair	n pos response of 1.00 (clearly correct)
`typical_of_concept`	30	0	0	-	12 (12)
`typical_of_property`	13	5	12	0.85	8 (8)
`implied_category`	25	2	3	0.63	9 (9)
`affording_activity`	22	5	3	0.66	9 (9)
`afforded_usual`	30	0	0	-	17 (17)
`afforded_unusual`	18	1	11	0.86	2 (2)
`variability_limited`	25	4	1	0.62	4 (4)
`variability_open`	24	6	0	-	6 (5)
`rare`	25	5	0	-	2 (2)
`unusual`	26	3	1	0.63	2 (1)
`impossible`	22	8	0	-	4 (4)
`creative`	24	6	0	-	2 (2)

Table 6.8: Overview of evaluation of random samples per relation.

6.3.2 Relation Distinctiveness

In addition to accuracy, I explore how well the crowd annotations can distinguish between different property-concept relations. Certain combinations of relations are unlikely (and should thus not occur frequently) while others are expected. For instance, *implied_category* is likely to occur in combination with `typical_of_concept`, as both relations describe close associations between a property and concept. In contrast, the relation `afforded_usual` should not occur at the same time as `afforded_unusual`, as the relations express radically different relationships between properties and concepts.

To analyze whether the crowd behaved as expected, I test whether property-concept pairs have been labeled with specific pairs of relations. When considering the example of the two typicality relations `typical_of_concept` and `typical_of_property`, I expect that almost all pairs annotated with `typical_of_property` have also been annotated with `typical_of_concept`. For example, *blood* is a typical example of things which are **red** (`typical_of_concept`), but **red** is also one of the first properties that come to mind when thinking of *blood*. However, only a small subset of pairs labeled with `typical_of_concept` should also be annotated with `typical_of_property`; green is a typical color of broccoli, but *broccoli* is most likely not the first thing that comes to mind when thinking of the color green. To test such assumptions, I calculate the proportion of pairs

annotated with relation (a) (e.g. `typical_of_concept` that are also annotated with relations (b) (e.g. `typical_of_property`).

Table 6.9 shows the results of this analysis for a number of selected pairs (the full results are included in Appendix). Only pairs for which both relations appeared in the the crowd annotation task are included. As expected, several relations that express a strong association between a property and all instances of a concept show high overlap. For instance, 83% of pairs annotated with `afforded_usual` are also annotated with `typical_of_concept` and 99% of pairs annotated with `typical_of_concept` are also annotated with `afforded_usual`. The fact that there is no complete overlap between any of the relations indicates that they still express distinguishable notions.

For some relations expressing strong, positive associations, it is expected that they should not have a high degree of overlap. While pairs annotated with `typical_of_property` should also be annotated with `typical_of_concept`, only a small subset of pairs annotated with `typical_of_property` should also be annotated with `typical_of_property`. Almost 60% of pairs annotated with `typical_of_concept` have also been annotated with `typical_of_property`. This intersection is relatively large and most likely indicates that many crowd workers most likely indicated that `typical_of_property` applied when they encountered a strong association between property and concept. For the other positive relations with expected distinctions, the intersections are much smaller. For `typical_of_concept` and `variability_limited`, a proportion of pairs assigned to both relations is expected; a typical property can still be variable (e.g. **red**-*apple*).

A third group of relations encompasses relations expressing positive associations and relations expressing negative relations that should be distinguished, but could possibly be confused by the crowd. Candidates for possible confusion are shown in the table. When considering `afforded_unusual` and negative relations, it can be observed that a relatively high proportion of pairs annotated with `afforded_unusual` were also annotated with `unusual` (37%) or `rare` (28%). This is not intended, but can be plausible as their relations can lead to similar interpretations by untrained annotators:

(15) a. All or most puppy(s) can swim/be used for swimming. This is not what they normally do or are used for. (`afforded_unusual`)

b. Usually, (a/an) puppy cannot swim/be used for swimming, but there could be a highly unusual situation in which (a/an) puppy can swim. (`unusual`)

The relations expressing variability also overlap with the negative relation `rare` for some pairs. This is to be expected, as people may have varying degrees of knowledge or different thresholds for viewing a variable property as a rare occurrence (e.g. **yellow**-*tomato*).

For negative relations, a relatively high degree of overlap is expected, as the relations mainly serve to facility the annotation process and already anticipate disagreement. As expected, the negative relations `rare` and `unusual` share a high degree of overlap. In contrast, the relations `rare` and `unusual` can clearly be distinguished from the relation `impossible`.

Overall, the analysis of overlap between properties indicates that the crowd can indeed make some fine-grained distinctions, but difficult semantic phenomena might not be reflected

accurately. The natural language statements translation expressing abstract semantic relations is not always easy to understand for untrained annotators. Furthermore, the statements might not be precise enough and leave room for interpretation.

exp.	rel1	rel1 with rel2	rel1 and rel2	rel2 with rel1	rel2
overlap pos.	afforded_usual	0.83	0.82	0.99	typical_of_concept
	afforded_usual	0.97	0.81	0.83	implied_category
	affording_activity	0.90	0.74	0.81	implied_category
	implied_category	0.85	0.74	0.85	typical_of_concept
	affording_activity	0.91	0.71	0.76	typical_of_concept
distinctions pos	typical_of_concept	0.59	0.58	0.98	typical_of_property
	typical_of_concept	0.47	0.32	0.51	variability_limited
	afforded_unusual	0.14	0.05	0.07	afforded_usual
	variability_limited	0.21	0.14	0.28	variability_open
distinctions pos-neg	afforded_unusual	0.37	0.26	0.46	unusual
	afforded_unusual	0.28	0.24	0.62	rare
	afforded_unusual	0.11	0.04	0.05	impossible
	rare	0.30	0.15	0.22	variability_open
	rare	0.34	0.14	0.20	variability_limited
overlap neg.	rare	0.88	0.61	0.66	unusual
	impossible	0.21	0.10	0.15	unusual
	impossible	0.05	0.03	0.05	rare

Table 6.9: Analysis of intersections between property-concept pairs annotated with relations. The table shows the proportion of pairs annotated with rel1 that have also been annotated with rel2 and vice-versa. The table also shows the proportion of pairs annotated with rel1 and rel1 out of all pairs annotated with either rel1 or rel2.

6.3.3 Discussion and conclusion

This section presented an evaluation of the crowd task on the level of specific property-concept relations. The analysis was based on 30 randomly selected property-concept pairs labeled with a relation based on the best performing filtering-and aggregation method introduced in Section 6.2.

The results of relation-accuracy indicate that overall, the crowd annotations can offer a relatively accurate reflection of different property-concept relations. However, it should be pointed out that the untrained annotators were not able to apply complex ideas correctly; the relations `typical_of_property` and `afforded_unusual` showed the highest numbers of incorrectly labeled pairs. On a more coarse-grained level, it could be observed that the pairs labeled with clearly positive relations contain positive example concepts of the property and pairs labeled with negative relations usually contain negative examples. This evaluation of relation accuracy has the limitation that only pairs labeled with a relation were

considered; it did not provide indications about false negatives (i.e. pairs that should have been labeled with a relation but were not).

In addition to relation accuracy, I considered the degree to which the crowd annotations reflect expected distinctions and expected overlap between property-concept relations. While the observed patterns mostly reflect that the crowd could, at least to some extend, make certain distinctions, the results also indicate that the relations `typical_of_property` and `afforded_unusual` may not always have been interpreted correctly.

The reason for the difficulty of the crowd to apply the two relations accurately may lie in the task setup or sentence formulation. It could be the case that the binary judgment of relatively complex sentences is not suitable for eliciting fine-grained distinctions from the crowd. In an alternative setup, annotators could see all possible relations at once and thus might be more likely to compare relations against each other. In addition, the formulation of the statements expressing the two relations could have been improved. Ultimately, it could be the case that the two relations in question are too abstract and complex for a crowd annotation task. In future research, it could be tested whether the two typicality relations could be distinguished in a psychologically-informed annotation setup.

6.4 Summary

In this chapter, I presented two strategies for evaluating crowd annotations collected in a complex semantic task: Section 6.2 presented an evaluation with respect to expected crowd behavior. It showed that crowd annotations largely follow expected patterns; disagreement occurs in instances that exhibit linguistic phenomena such as vagueness or ambiguity or instances which require specific knowledge. Disagreement can thus be seen as a valuable signal that has the potential to highlight specific phenomena in the data. Furthermore, the evaluation showed that a task-inherent coherence-check offered the best method of detecting unreliable annotations and filtering judgments.

The second evaluation strategy (Section 6.3) focused on label accuracy of specific, fine-grained property-concept relations. The results indicate that accuracy differs between relations. While many relations revealed a high proportion of accurate judgments, two showed less reliable results. In particular, annotators seem to have problems with abstract and complex semantic relations that require fine-grained distinctions. On a more coarse-grained level, the judgments reflect an accurate classification of concepts as positive or negative examples of a property.

Based on both evaluations, it can be concluded that the crowd annotation task was partially successful; it resulted in a set of relatively reliable positive and negative examples of semantic properties. In addition, a number of semantic relations have been annotated reliably. Two relations, however, have triggered confusions between fine-grained semantic notions. This should not be seen as a reason not to use the dataset; rather, the limitations should be considered in further analysis.

7. A Corpus of Properties, Concepts, and Relations

7.1 Introduction

This chapter presents an analysis of the diagnostic dataset of properties, concepts, and relations between properties and concepts. The main purpose of the dataset is to study the semantic properties captured by context-free embedding models. For example, the dataset should reveal whether embedding representations carry semantic information about the fact that lemons and sunflowers are usually yellow, but limes and violets are not or that seagulls and airplanes can fly, but penguins and cars cannot. Beyond this, the dataset should have the potential to give insights into potential mechanisms that determine whether specific semantic information is reflected by distributional co-occurrence patterns extracted from corpora and underlying the embedding models. In this chapter, I investigate to what extent the dataset constitutes a suitable diagnostic resource and thus address **step 2c**:[1]

Step 2-c Assess the resulting diagnostic dataset in terms of its ability to yield insights about specific hypotheses and its adherence to methodological requirements.

To fulfil its diagnostic purpose, the dataset has to adhere to the following criteria: Firstly, the dataset has to contain challenging and informative examples for diagnostic experiments (see Chapter 4). Diagnostic experiments require positive and negative examples of a particular piece of information. In the case of semantic properties, the target information is a particular property, while the positive and negative examples are the embedding representations of example concepts. A diagnostic classifier should provide insights into whether the embedding representations capture the target property; if the classifier can learn to distinguish positive from negative examples, it indicates that the property is indeed captured. The outcome of a diagnostic classification experiment can only be valid if the property information is the only information by which positive examples can be distinguished from negative examples. One focus of the analysis in this chapter is to establish to what degree this is indeed the case for the diagnostic dataset.

Secondly, the dataset should be able to provide insights that go beyond establishing whether a certain property is captured by embeddings or not and provide insights into potential underlying mechanisms that govern whether information tends to be captured by distributional data or not (see Chapter 3). To investigate such mechanisms, properties and concepts are linked with relations that should allow for testing specific hypotheses grounded in theoretical and empirical research. For instance, it could be expected that embedding representations are

[1] The full dataset and code used for analysis can be downloaded from this repository: `https://github.com/PiaSommerauer/PropertyConceptRelations`.

good at capturing semantic information about afforded and usually performed activities (e.g. **used_in_cooking**: *pasta*), but lack information about afforded, but usually not performed activities (**roll**: *candle*). The second goal of this chapter is to establish to what extent the dataset is indeed suitable for testing these hypotheses.

The results of the analysis provides a fine-grained overview of potential risks when drawing conclusions from diagnostic experiments. The 21 property datasets that make up the diagnostic dataset each have different profiles; while twelve out of the 21 property datasets can be considered low risk, nine property datasets run risk of yielding misleading conclusions. They can still be used in diagnostic experiments, but their results should be considered in the light of their potential limitations.

The analysis of property-concept relations indicates different patterns for different property types. When considering their explanatory power with respect to the hypotheses outlined in Chapter 3, however, the analysis reveals complex interactions between different property concept relations. Isolating individual relations to test their effect is hardly possible.

Despite its limitations, the dataset goes beyond existing diagnostic resources by making the following contributions:

- Each property dataset contains verified positive and negative examples.
- The dataset has enough examples to allow for training and testing on a held-out test set.
- The dataset comes with a fine-grained analysis of potential limitations on the level of individual properties.

Beyond these contributions for diagnostic experiments, the dataset constitutes a rich resource of fine-grained aspects of common sense knowledge. Even though the property-concept relations are not yet suitable for testing specific hypotheses, they offer a detailed characterization of how properties can relate to concepts.

The remainder of this chapter is structured as follows: After an outline of the post-processing steps used to curate the dataset (Section 7.2), Section 7.3 provides an overview of central components of the dataset. Section 7.4 presents an analysis of the property datasets with respect to their suitability for diagnostic experiments and explainable power. Section 7.5 presents 'property-profiles' that summarize the suitability of property datasets for diagnostic experiments.

7.2 Post-Processing

This section presents the post-processing steps undertaken to remove noisy annotations and assign property-concept pairs to fine-grained and coarse-grained property-concept relations, as well as binary classes for diagnostic experiments.

Annotation task The post-processing steps are closely tied to the annotation task set-up and procedure (described in detail in Chapter 5): The goal of the task was to annotate concepts (e.g. *lemon*, *chocolate*) in terms of whether they are examples of a semantic property (e.g. yellow). Specifically, property-concept pairs should be labeled with fine-grained relations

that connect them. To accomplish this, property-concept-relation triples were translated into natural language statements, such as the following example:

(16) "Blood" is one of the first things which come to mind when I hear "red' because (a/an) blood is a typical example of things which are red'. (triple: **red**-*blood*-`typical_of_property`)

Crowd annotators had to indicate whether they agreed or disagreed with the statement. In a single annotation batch (which should take around seven minutes), crowd annotators saw all possible relations for a given property-concept pair. A single batch always contained multiple property-concept pairs to ensure diversity.

Filtering Before assigning labels or relations, all annotations were filtered following the method outlined in the previous chapter (Chapter 6). The core principle of the filtering method is the removal of annotations by annotators whose answers contain logical contradictions within an annotation batch. Contradictions are defined as property-concept pairs that have been annotated with mutually exclusive property-concept relations. In particular, relations that connect a property to all or most instances of a concept (MOST-ALL category) and relations that connect properties to few or no instances of a concept (FEW-NONE category) cannot apply to the same property-concept pair. Consider the following two annotation units expressing property-concept-relation triples as statements:

(17) a. "Wine" is one of the first things which come to mind when I hear "blue' because (a/an) wine is a typical example of things which are blue'. (triple: **blue**-*wine*-`typical_of_property`)

b. Usually, (a/an) wine is not blue, but there could be a highly unusual situation in which (a/an) wine is blue. (triple: **blue**-*wine*-`unusual`)

Statement 17a implies that all instances of *wine* are **blue**, while statement 17b implies that only very few instances of *wine* are **blue**. If a worker agreed with both statements in the same annotation task, this was counted as a contradiction, as both statements cannot be true at the same time. In principle, all annotations from annotators who gave contradictory answers were removed from the annotation batch (i.e. statements annotated in a single annotation session).

Not all instances in the annotation task were as clear as Example 17a and Example 17b. Several statements contained ambiguity and vagueness. In such cases, contradictions can be justified. If many annotators submitted annotations that contained contradictions, it can be assumed that the statements in the annotation batch contained difficult instances that justified the contradictions. Therefore, not all submissions containing contradictions were removed. Instead, the threshold for the 'accepted' number of contradictions was adjusted based on the behavior of all annotators who worked on a batch.

Fine-grained relation-assignment The goal of the annotation task was to assign relations to property-concept pairs. To decide whether a relation should be assigned to a property-concept pair, I use a majority-vote of all annotations left after filtering: If more than 50% of annotators agreed with a statement, the relation is assigned to the property-concept pair.

Coarse-grained relation assignment The fine-grained relations fall into categories depending on the subset of instances of a concept they apply to: MOST-ALL, SOME, FEW-NONE. I assign the coarse-grained relation based on the fine-grained relation with the highest positive response rate (i.e. proportion of annotators who agreed with a statement). This strategy can result in multiple coarse-grained relations per pair if multiple relations have the same positive response rate. For instance, the pair **yellow**-*marigold* has a positive response rate of 1.0 for the relations `typical_of_concept`, `typical_of_property`, and `variability_limited`. Therefore, it receives the coarse-grained relations MOST-ALL and SOME. Pairs whose relations fall into mutually exclusive coarse-grained relations categories (MOST-ALL and FEW-NONE) are not assigned to a coarse grained relation and not included in further analysis.

Binary labels To make use of the data in diagnostic experiments, it is necessary to assign binary labels to each concept in a property dataset. I draw the line between positive and negative examples between the coarse-grained relations SOME and FEW-NONE. I also include concepts which feature both the coarse-grained relation SOME and the coarse-grained relation FEW-NONE in the positive class. Such examples may run risk of containing noise, but overall seem to contain justified positive examples (e.g. **red**-*melon*, **black**-*crocodile*). The examples listed here can reasonably be described as positive examples of their properties, but may trigger slightly different interpretations among annotators. Consequently, the negative class consists of concepts whose relations only fall into the FEW-NONE category.

7.3 Dataset Overview

In this section, I provide an overview of the diagnostic dataset. I provide general statistics and consider the core components of the dataset: properties, concepts, and their relations.

General statistics Table 7.1 presents an overview of the entire dataset. In total, the diagnostic set encompasses 21 properties and 1756 different concepts. Each concept can be part of one or more property sets. In total, this results in 3304 property-concept pairs. The property concept pairs can be linked by 12 different fine-grained relations. A combination of property, concept, and fine-grained relation makes up an annotation unit. In the annotation task, annotators judged individual annotation units (expressed as natural language statements) one by one. In total, the dataset encompasses 30650 annotation units. The fine-grained relations can be categorized into 5 different coarse-grained relations (e.g. ALL, ALL-SOME).

The dataset has been filtered and post-processed as outlined in Section 7.2. Most importantly, the post-processing steps removed low-quality annotations. The results of this filtering step can be observed in Table 7.2: Overall, the mean number of annotations per unit was reduced to 8 (compared to 10). The overall inter-annotator agreement (measured by Krippendorff's alpha) rose from 0.36 to 0.40. The mean time spent per annotation unit and annotator was about 9 seconds. Post-processing resulted in a higher number of pairs that could not be assigned to a label.

	total n
units	30650
pairs	3304
concepts	1756
properties	21
fine-grained relations	12
coarse-grained relations	5

Table 7.1: Overview of the diagnostic dataset.

	raw	clean
mean annotations per unit	10.08	8.06
Krip. alpha	0.36	0.40
mean duration per unit	9.24	9.25
pairs no label	267	290

Table 7.2: Effect of post-processing.

	properties
perceptual	juicy black square blue cold yellow green round sweet red warm hot
activities	fly roll swim lay_eggs
complex	used_in_cooking dangerous
parts/material	wings wheels made_of_wood

Table 7.3: Overview of properties and property types.

Properties The core of the diagnostic dataset consists of 21 semantic properties. The properties can be divided into four rough categories, as shown in Table 7.3. One goal of the dataset is to have a verified selection of positive and negative examples for each property. Ideally, the positive and negative classes should be balanced. Table 7.4 shows the mean number of positive and negative examples per property. Overall, properties tend to have more positive than negative examples (86 positive compared to 57 negative examples). Around 14 examples per property could not be assigned to a class because they contained contradictory annotations (see Section 7.2). The positive and negative classes are not completely balanced, but contain a substantial number of examples. As shown in Chapter 8, the property-datasets are large enough for diagnostic classification experiments which involve training a classifier and testing on a held-out test set.

	mean
examples pos	86.10
examples neg	57.43
examples no label	13.81

Table 7.4: Mean number of examples per semantic property.

Concepts Each concept in the diagnostic dataset is part of one or more property datasets. The distribution of concepts over property datasets and their positive and negative classes is summarized in Table 7.5. On average, each concept is part of nearly two property datasets.

	mean
properties	1.88
positive examples	1.03
negative examples	0.69
invalid examples	0.17

Table 7.5: Overview of distribution over property datasets and classes on the level of individual concepts (mean).

coarse-grained	relation	alpha	seconds	pairs	properties	candidate pairs
most-all	implied_category	0.53	7.73	1060	21	3045
	typical_of_concept	0.55	6.84	1056	21	3045
	typical_of_property	0.42	7.23	633	21	3045
	afforded_unusual	0.27	9.40	107	4	623
	afforded_usual	0.63	7.81	214	4	623
	affording_activity	0.51	7.22	732	17	2422
some	variability_limited	0.34	9.32	967	21	3045
	variability_open	0.32	9.05	672	17	2422
few-none	rare	0.28	7.07	630	21	3095
	unusual	0.30	8.11	839	21	3095
	impossible	0.41	8.04	617	21	3095
	creative	0.17	7.55	400	21	3095

Table 7.6: Overview of inter-annotator agreement (measured by Kippendorff's alpha), duration, and distribution over pairs and properties on the level of fine-grained relations.

Relations The third central element of the dataset are fine-grained property-concept relations. The relations reflect hypotheses about underlying factors that impact whether property-evidence is likely to be encoded in distributional data. For instance, highly implied information that is likely to apply to a larger semantic category (e.g. **round**-*lemon*) is not expected to be mentioned explicitly and systematically in corpus data. This is reflected by the relation `implied_category`.

The property-concept relations fall into three coarse-grained categories. The coarse-grained categories are based on the number of instances of a concept the property applies to (MOST-ALL, SOME, or FEW-NONE). Table 7.6 provides an overview of all relations sorted by coarse-grained category. The table provides information about annotator behavior (inter-annotator agreement measured by Krippendorff's alpha and annotation time), and the distribution of property-concept pairs over relations.

Overall, agreement in the MOST-ALL category tends to be higher than in the SOME category. The inter-annotator agreement for negative relations and for the relation that

Relation combination	pairs	properties
impossible	313	19
rare unusual	242	19
affording_activity implied_category typical_of_concept typical_of_property variability_limited	179	12
variability_open	175	12
creative impossible	170	18
affording_activity implied_category typical_of_concept typical_of_property	155	13
variability_limited	114	15
affording_activity implied_category typical_of_concept variability_limited	100	11
rare unusual variability_limited	93	13
afforded_usual implied_category typical_of_concept typical_of_property	88	4

Table 7.7: Top 10 relation configurations.

expresses a creative link between property and concept is lowest. This is to be expected, as people tend to have different thresholds for when they call a property-concept combination rare, unusual, or impossible and for what combinations they could understand in a figurative manner.

When considering the distribution of properties over relations, it should be kept in mind that not all relations can apply to all properties; for instance, only activity-properties can have the relations `afforded_usual` and `afforded_unusual`. When considering the distribution of property-concept pairs over relations, it should be noted that each pair can be assigned to multiple relations. For example, the property concept pair **red**-*apple* is likely to be annotated with the relations `variability_limited` and `typical_of_concept`. This results in a high number of pairs per relation and high overlap between pairs assigned to different relations. To illustrate the effect of this distribution, the top ten relation configurations are shown in Table 7.7. It can be observed that combinations of many different relations from the MOST-ALL and SOME categories tend to be frequent. Such combinations of many positive relations complicate the analysis of individual relations in diagnostic experiments, as it is hardly possible to distinguish the effects of individual relations.

Coarse-grained relations, in contrast, are distributed in a more straight-forward manner; each property-concept pair is assigned to a single coarse-grained relation. The distribution of properties and pairs over coarse-grained relations as well as the annotation behavior on the level of coarse-grained relations are shown in Table 7.8. The two extreme ends of the spectrum (MOST-ALL and FEW-NONE) appear in all property datasets. The relations linking properties to a subset of concept instances (e.g. SOME: **red**: *apple*) do not appear in all property datasets. This is plausible, as some properties are unlikely to apply to only subsets rather than all or no instances (e.g. **wings**). The inter-annotator agreement is highest for clear positive relations (0.53 for ALL-SOME and lowest for relations that link properties to a subset of concept instances (0.29 for SOME, SOME-FEW).

relation	alpha	seconds	pairs	properties	candidate pairs
all	0.53	7.87	3092	21	7996
all-some	0.49	8.28	827	19	1936
some	0.29	9.32	1475	20	6690
few-some	0.29	6.99	194	17	710
few	0.36	7.66	2289	21	10908

Table 7.8: Overview of inter-annotator agreement (Kippendorff's alpha), duration, and distribution over pairs and properties on the level of coarse-grained relations.

Summary This section has presented an overview of the diagnostic dataset in terms of its three main components: properties, concepts, and relations. Overall, the statistics presented in this first analysis indicate that the dataset is indeed suitable for diagnostic experiments, as it provides decently sized property datsets containing verified positive and negative examples.

In addition to being used in diagnostic experiments, the dataset aims to provide a basis for testing hypotheses about the underlying dynamics that govern whether property-information tends to be expressed explicitly in distributional data. The hypotheses are tied to fine-grained property-concept relations. The analysis of these relations provides first indications that the interactions between the relations are complex. Multiple relations can apply to the same property-concept pair. The relations form complex configurations that do not necessarily allow for analyzing the effect of individual relations.

Despite the potentially limiting complexity of property-concept relations, the dataset can still constitute a diagnostic tool for semantic properties. The subsequent section presents an analysis of the 21 property datasets. In particular, focus is placed on the degree to which individual property datasets can provide valid results in diagnostic experiments and to what degree they can provide explanatory insights.

7.4 Analysis of Property Datasets

In this section, I provide a characterization of the property datasets with respect to factors that may impact the suitability of the datasets for diagnostic experiments and their explanatory power.

The main criteria for valid diagnostic classification experiments are the following: (1) The property has to be learnable given the dataset. (2) The property should be the only factor that allows to distinguish positive from negative examples. To establish whether the datasets fulfil these criteria, I consider the dataset sizes and class distribution as well as inter-annotator agreement on property level (Section 7.4.1). As a next step, I assess the degree to which positive and negative examples can be separated without necessarily having information about the target property (Section 7.4.2). This is followed by an analysis of different lexical features (first and foremost frequency and ambiguity) in the datasets (Section 7.4.3). Both features can impact embedding vectors and thus the classification task. Finally, I provide a characterization of property datasets in terms of the fine-grained relations represented by them (Section 7.4.4).

prop	#concepts	pos	neg	no-label	total-valid	prop-pos	alpha	seconds
used_in_cooking	180	107	68	5	175	0.61	0.52	7.93
wings	180	81	82	17	163	0.50	0.52	7.17
fly	180	65	106	9	171	0.38	0.43	8.09
green	180	95	70	15	165	0.58	0.42	6.38
hot	153	104	46	3	150	0.69	0.41	7.31
cold	110	73	23	14	96	**0.76**	0.40	6.86
swim	180	105	44	31	149	0.70	0.40	7.97
lay_eggs	153	77	68	8	145	0.53	0.40	8.65
wheels	114	75	29	10	104	**0.72**	0.38	8.32
blue	180	60	110	10	170	0.35	0.36	6.94
sweet	173	101	61	11	162	0.62	0.36	8.01
juicy	180	92	62	26	154	0.60	0.34	6.95
dangerous	140	78	57	5	135	0.58	0.34	8.54
warm	180	133	38	9	171	**0.78**	0.34	6.93
made_of_wood	151	103	42	6	145	**0.71**	0.33	8.51
black	151	89	55	7	144	0.62	0.31	7.42
yellow	174	44	88	42	132	0.33	0.27	7.72
square	119	90	21	8	111	**0.81**	0.26	7.60
round	137	107	23	7	130	**0.82**	**0.23**	9.29
red	169	95	68	6	163	0.58	**0.23**	10.48
roll	120	62	40	18	102	0.61	**0.15**	9.14

Table 7.9: Overview of property sets in terms of total number of concepts per property (#concepts), number of positive (pos) and negative (neg) examples, number concepts without valid labels (no-label) and the total number of examples with a valid label (total-valid). In addition, the proportion of positive examples (prop-pos), Kippendorff's alpha (alpha), and the mean duration per judgment (seconds) are shown.

7.4.1 Class Distribution and Agreement

Fundamental criteria for good diagnostic property sets are enough positive and negative examples and a low chance of noise. As a first step in assessing the quality of the datasets, I consider the distribution of positive and negative examples and the inter-annotator agreement per property dataset.

Table 7.9 lists the sizes of the positive and negative class together with the inter-annotator agreement after pre-processing ('alpha'). The degree of class imbalance is measured by the proportion of positive examples out of all valid examples (indicated by 'prop-pos'). In a balanced dataset, this proportion would be 0.5.

Class balance All datasets with a class that makes up more than 70% of the examples are marked in bold. This high degree of imbalance applies to six out of 21 properties: **cold**, **wheels**, **warm**, **made_of_wood**, **square**, and **round**. Overall, most property sets tend to be skewed towards the positive class (with the exception of **yellow** and **blue**). For datasets with high class imbalance, learning to identify the target property is more difficult than for datasets with balanced distributions. A reason for this class imbalance is that the dataset extension

	wiki_corpus			giga_corpus			googlenews		
	f1	p	r	f1	p	r	f1	p	r
wings	**0.85**	0.85	0.85	**0.78**	0.80	0.78	**0.81**	0.83	0.81
used_in_cooking	**0.80**	0.86	0.80	**0.88**	0.91	0.87	**0.88**	0.90	0.87
round	**0.73**	0.79	0.69	0.58	0.78	0.51	0.60	0.70	0.55
lay_eggs	0.67	0.81	0.70	0.69	0.74	0.72	0.74	0.74	0.74
square	0.62	0.83	0.58	0.72	0.84	0.69	0.67	0.84	0.63
red	0.62	0.63	0.62	0.58	0.59	0.58	0.54	0.57	0.55
fly	0.61	0.62	0.63	0.54	0.58	0.52	0.51	0.61	0.51
warm	0.59	0.68	0.55	0.56	0.64	0.51	0.65	0.69	0.62
made_of_wood	0.57	0.58	0.56	0.58	0.54	0.66	0.59	0.60	0.58
sweet	0.57	0.57	0.57	0.53	0.53	0.56	0.63	0.63	0.63
wheels	0.56	0.54	0.57	**0.78**	0.79	0.80	**0.80**	0.86	0.83
swim	0.55	0.75	0.56	0.54	0.75	0.55	0.51	0.82	0.53
blue	0.53	0.67	0.54	0.50	0.71	0.53	0.50	0.61	0.50
dangerous	0.53	0.60	0.59	0.52	0.55	0.57	0.53	0.58	0.59
juicy	0.53	0.54	0.52	0.54	0.55	0.54	0.64	0.64	0.65
yellow	0.52	0.55	0.50	0.51	0.55	0.50	0.52	0.55	0.51
hot	0.52	0.54	0.51	0.52	0.63	0.50	0.53	0.63	0.51
black	0.52	0.51	0.56	0.71	0.72	0.73	0.59	0.69	0.59
cold	0.51	0.53	0.50	0.53	0.55	0.52	0.58	0.73	0.56
green	0.49	0.55	0.50	0.54	0.57	0.54	0.62	0.63	0.62
roll	0.47	0.47	0.51	0.46	0.37	0.60	0.49	0.55	0.51

Table 7.10: Overview of k-means clustering analysis ($k = 2$): Performance is assessed in terms of precision, recall, and f1-score (weighted). A high f1-score indicates that the positive and negative class are easily separable.

strategy (see Chapter 4 for details) used to find challenging examples in an embedding model favored positive examples. One goal of this strategy was to collect challenging negative examples that have a high similarity to already collected positive examples (as well as positive examples). It is likely that the strategy resulted in more positive than negative examples.

Inter-annotator agreement Properties with the top three lowest agreement values are marked in bold. The three properties with the lowest agreement values are **roll**, **red**, and **round**. A possible reason for this low agreement is that the three properties in question were annotated at the beginning of the annotation process (for details, see Chapter 5). At this stage, the crowd workers were still new to the task. At later stages in the annotation process, participation in the task was only open to crowd workers who had delivered reliable work in previous annotation batches.

7.4.2 Class separability

Another aspect that is crucial for the validity of diagnostic classification experiments is that positive and negative examples of a property should only be distinguishable by the target property. While it is impossible to ensure that this criterium holds for every single example pair, the dataset was designed in such a way that positive and negative examples should have

	googlenews			giga_corpus			wiki_corpus		
	closest neg	furthest pos	sp. rho	closest neg	furthest pos	sp. rho	closest neg	furthest pos	sp. rho
used_in_cooking	86	168	**0.78**	98	151	**0.80**	70.0	155.0	**0.79**
wings	36	138	**0.77**	15	126	**0.67**	45.0	110.0	**0.80**
lay_eggs	13	92	**0.65**	6	73	0.65	47.0	115.0	**0.75**
dangerous	15	127	0.64	20	108	**0.68**	32.0	136.0	0.68
square	35	117	0.56	51	115	0.55	9	115	0.52
wheels	30	110	0.54	31	104	0.47	19	112	0.47
fly	18	157	0.50	5	124	0.54	34	168	0.59
cold	24	108	0.44	26	107	0.43	22	109	0.46
swim	31	165	0.41	9	147	0.48	13	178	0.38
juicy	9	171	0.38	14	171	0.21	3	174	0.22
green	2	172	0.37	1	169	0.29	1	176	0.28
hot	20	151	0.36	17	144	0.34	18	148	0.38
sweet	13	169	0.28	8	161	0.22	7	167	0.29
warm	14	176	0.19	10	169	0.24	42	179	0.26
yellow	3	154	0.16	1	140	0.18	2	152	0.27
made_of_wood	10	141	0.15	2	115	0.19	2	147	0.18
roll	12	117	0.14	1	102	0.14	7	119	0.26
black	3	139	0.11	3	131	0.06	3	149	0.12
red	8	160	0.03	5	156	-0.01	2	164	-0.06
blue	5	177	0.01	3	170	0.13	10	177	0.22
round	12	135	-0.06	18	130	-0.01	15	135	0.01

Table 7.11: Overview of separability measured by cosine distance from centroid and Spearman Rho (sp. rho) correlation between labels and distance from centroid. A high correlation indicates that negative examples have a higher distance to the centroid than positive examples.

a challenging distribution. In practice, positive examples should be semantically diverse and located in different areas of an embedding space model. Negative examples should be similar to positive examples not easily separable from them in terms of general dissimilarity.

In this section, I assess the risk of easily separable classes based on general vector similarity. To address this issue, I use two strategies to test how easily positive examples can be separated from negative examples based on their location in the embedding space: (1) I use k-means clustering to test to what degree examples 'naturally' fall into the positive and negative class. For this purpose, I use two possible clusters ($k = 2$). (2) I use distance to the centroid vector of the positive class to assess to what extent positive and negative examples can be separated in terms of distance to the centroid of the positive class. This strategy assesses to what degree negative examples are dissimilar to positive examples, regardless of whether they form a cluster themselves. Both analyses are conducted for three Word2Vec skip-gram embedding models: a model trained on the Wikipedia corpus (2017 dump), a model trained on the Gigawords words corpus, and the Googlenews Word2vec model (henceforth wiki, giga, and google). The wiki and giga models were trained following the settings recommended by Levy and Goldberg (2014).[2] I chose these models as they are used in the diagnostic experiments presented in Chapter 8. The wiki and giga corpus are used for corpus analysis in Chapter 9.

Clustering The results of the clustering analysis are shown in Table 7.10. The performance of clustering is measured by means of comparing the clusters to the actual labels (measured by precision, recall, and weighted f1-score). For most properties, unsupervised clustering of vectors yields low performance (most properties yield f1 scores around 0.55). This indicates that overall, the examples do follow a challenging distribution, as they do not seem to naturally fall into the positive and negative class.

While no property yields a perfect score, it can be seen that the two property sets **wings** and **used_in_cooking** are among the properties with the highest clustering performance sets in all three models (marked in bold). It should be noted that high performance indicates a high degree of separability, which means that the property datasets may not be challenging in a diagnostic experiment. High clustering performance is not desirable. The property **wheels** has a high degree of separability in the giga and google models, while **round** is highly separable in the wiki model. For the two part-properties **wings** and **wheels**, this relatively high degree of separability is to be expected: Both properties tend to apply to traditional, possibly fine-grained, semantic categories (**wings**: BIRD, WINGED INSECT, FLYING VEHICLE; **wheels**: WHEELED VEHICLES). Likewise, the function property **used_in_cooking** is likely to apply to the relatively coherent categories of FOOD and KITCHEN APPLIANCES.

Centroid The results of the centroid analysis are shown in Table 7.11. The table shows the correlation (measured by Spearman Rho) between distance from the centroid calculated over the positive class and positive and negative labels. A high correlation indicates that all positive

[2]The wiki and giga model can be found in this repository: `https://bitbucket.org/PiaSommerauer/distributionalmodels`. The google model can be downloaded from `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?resourcekey=0-wjGZdNAUop6WykTtMip30g`

examples tend to be close to the centroid, while all negative examples tend to be further away. In a scenario in which the positive examples are highly diverse and negative examples are similar to positive examples, this would not necessarily be the case. Overall, the correlations between labels and distance to centroid are indeed low: most properties score below 0.5. While this indicates that for most properties, there is a positive correlation between class distribution and distance to the centroid, the property sets still contain challenging examples.

In addition to the correlation, the table also shows the ranks of the negative example with the smallest distance to the centroid and the positive example with the greatest distance to the centroid. These ranks can be seen as an indication of whether the examples follow a challenging distribution: A low rank for the closest negative example indicates that a negative example is located close to positive examples. Vice-versa, a high rank for the furthest positive example indicates that a positive example is located far away from other positive examples. It can be observed that property datasets with low correlations also tend to follow such a challenging distribution. For example, in the Googlenews model the property **blue**, which has a correlation of 0.01, the closest negative example appears on rank 5, while the furthest negative example appears on rank 177. Similar extremes can be observed for **red** and **black**. In comparison, for properties with a high correlation (e.g. **used_in_cooking**), the closest negative example does not appear until rank 86 in Googlenews. Overall, the furthest positive examples tend to appear on a challenging (i.e. high) rank (e.g. 168 for the property **used_in_cooking** in Googlenews).

Across all three models, the properties **used_in_cooking** and **wings** are among the properties with the three highest correlations. In addition, the property set for **lay_eggs** shows a high correlation in google and wiki. The property set for **dangerous** shows a high correlation in the giga corpus. These high correlation scores indicate that the property datasets contain a comparatively low number of challenging examples and run risk of allowing for successful classification without identifying the target information.

Both analysis methods also yield property sets that are likely to be challenging: The clustering analysis shows low performance for **roll** across all models. In google, **green** and **cold** also show low performance. In wiki and giga, **yellow** and **blue**, **swim** and **fly** also show low performance and are thus likely to follow a challenging distribution. The lowest correlations and most challenging distributions are shown by the color properties **blue**, **red**, and the shape property **round**.

7.4.3 Distribution of Lexical Features

A major risk of diagnostic classification is that embedding representations can be classified correctly without identifying the target information. A high degree of class separability (e.g. measured by unsupervised clustering performance) can provide a first indication of this risk. In a second step, I consider obvious lexical features that may correlate with one of the classes and cause the high separability. Beyond the risk of correlation, lexical features may also increase the difficulty of detecting information in an embedding representation. Two factors that are likely to impact the embedding representations of words are frequency (Dubossarsky et al., 2017) and ambiguity (Del Tredici and Bel, 2015). In addition to frequency and ambiguity, words can be characterized by a number of other features. An extensive collection of different

psycholinguistic features (e.g. concreteness of a concept) is provided by the MRC database (Brysbaert et al., 2018). I use the information recorded in the database to characterize each property dataset.

Frequency Low frequency in corpus data means that an embedding model has seen few examples of a word. As shown by Sahlgren and Lenci (2016) and discussed in Chapter 2 low frequency is likely to yield embeddings of lower quality. Low frequency words thus run risk of containing little information. If all words in a positive or negative class have a low corpus frequency, this may introduce an unwanted correlation.

To test the frequency distributions, I consider the frequencies of concepts in the wiki and giga corpus[3] The frequency distributions over the positive and negative class in each property dataset are shown in Figure 7.1 for the giga corpus and Figure 7.2 for the wiki corpus. The box plots show the distribution of the frequencies on a logarithmic scale as follows: The boxes and whiskers show four quartiles of the distribution. The bars in the boxes indicate the median. Outliers are shown as dots.

In the Gigawords corpus (Figure 7.1), it can be observed that the property datasets for **lay_eggs** and **swim** tend to have more low frequency words than the other datasets. The datasets for **square** and **used_in_cooking** are characterized by a striking difference between the positive and negative class, which could constitute an unwanted bias. In the Wikipedia corpus (Figure 7.2), the following observations can be made: Overall, the datasets for **lay_-eggs**, **swim** and **used_in_cooking** tend to have more low-frequency words than the other datasets. In addition, **lay_eggs, roll**, and **square** are also characterized by a stark difference between the positive and negative class, which may interfere with diagnostic experiments.

Ambiguity Ambiguity can also be reflected by embeddings and may lead to specific features in embedding vectors. Words with multiple senses tend to be located between the lexical fields reflecting the senses, depending on their distribution in the underlying corpus data (Del Tredici and Bel, 2015). If most words from a property dataset are particularly ambiguous, the property is likely to be more difficult to detect.

To approximate the degree of ambiguity of a word, I use the number of synsets it has in the Princeton WordNet hierarchy (Fellbaum, 2010; Miller, 1995). In this resource, synsets represent groups of words which can be used synonymously. Words with several senses appear in multiple synsets. The number of synsets a word appears in can be used as a proxy for its degree of ambiguity. It should, however, be noted that the synsets in Princeton WordNet represent rather fine-grained distinctions. This is by no means an exhaustive characterization of ambiguity, but it provides some indication of ambiguity in the property datasets. The distribution of synsets in the property datasets is depicted in Figure 7.3. It can be seen that the datasets are relatively comparable in terms of the number of synsets associated with a word. Overall, the dataset for **made_of_wood** seems to show a high degree of ambiguity.

[3]The same corpora are used to train models for diagnostic experiments (Chapter 8) and for corpus analysis (Chapter 9).

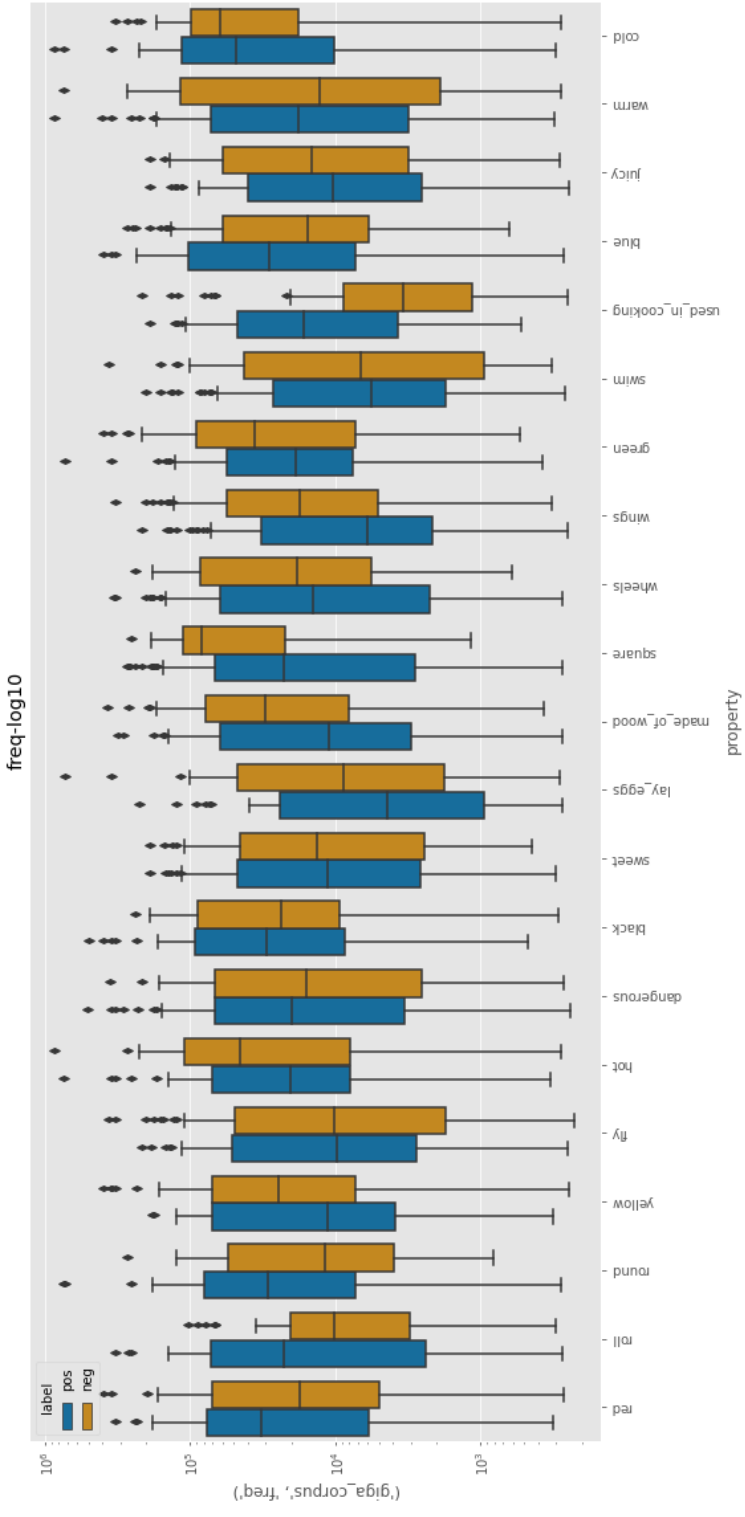


Figure 7.1: Frequency giga.

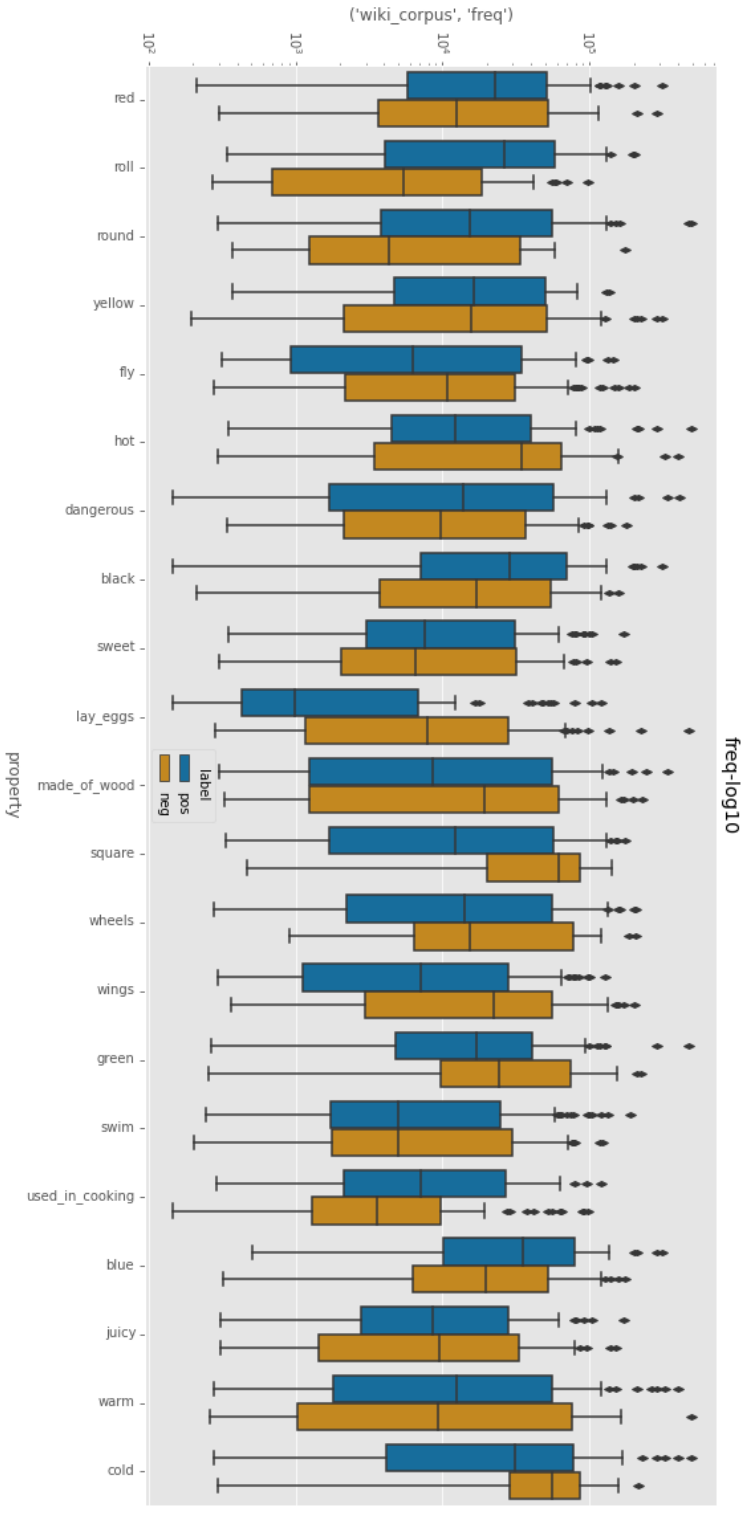


Figure 7.2: Frequency wiki.

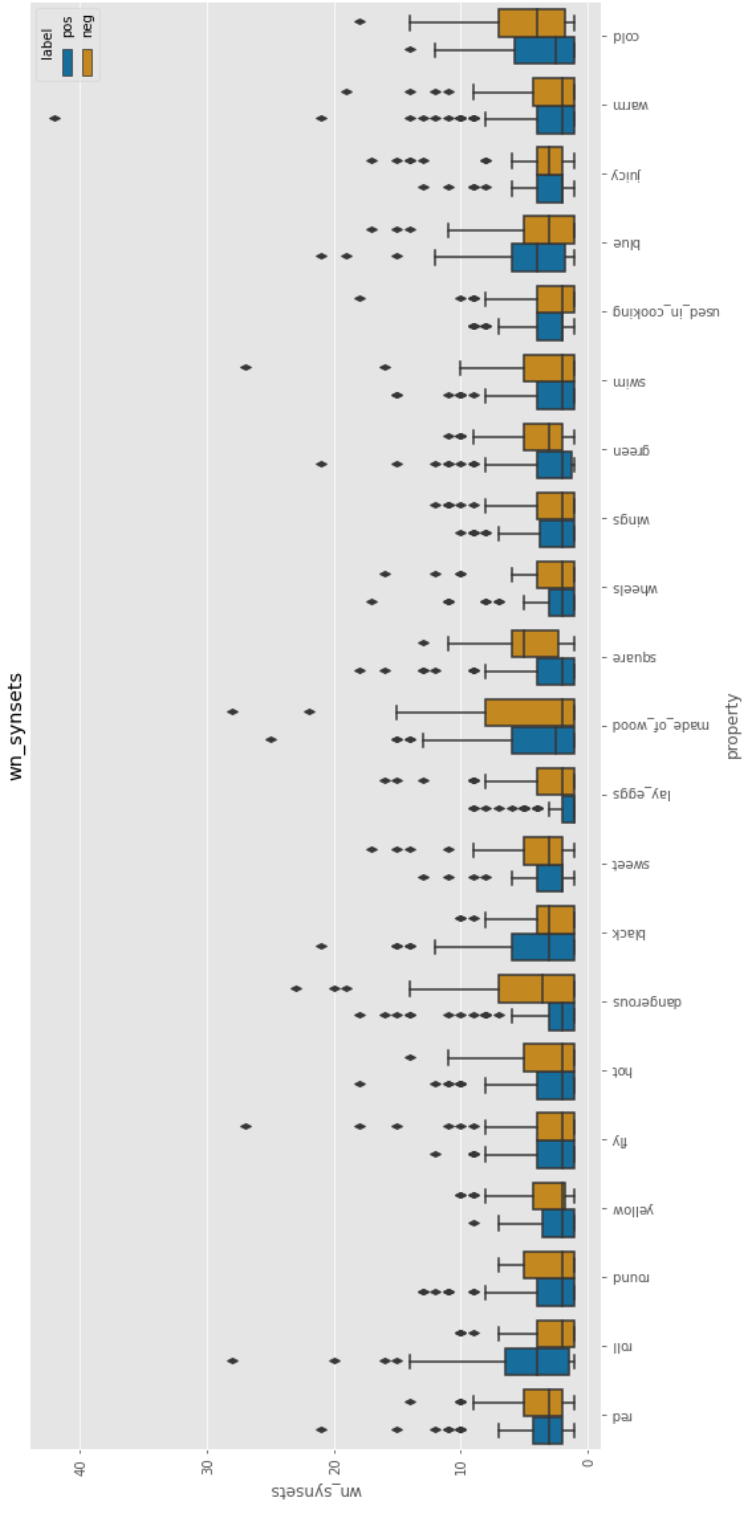


Figure 7.3: Number of wn synsets.

Psycholinguistic Features In addition to frequency and ambiguity, it is possible to measure the distribution of psycholinguistic features recorded in the MRC database in the property datasets. The database contains ratings for various psycholinguistic features (such as concreteness) on a lexical level. I consider features that may have an impact on annotation behavior as well as the distribution of words in the semantic space: concreteness, familiarity, and imageability. For all three features, it is possible that annotators find it easier to judge words that have a high rating. The potential impact on an embedding model is less straight-forward. One possibility is that words with high concreteness scores tend to be placed in a different area of the space than words with more abstract meaning. Imageability is not the same as concreteness, but is likely to be correlated with it to some degree. Another potential impact may be a correlation with frequency: While familiarity cannot be translated to frequency, it is likely to correlate with it. The ratings in the MRC database do not cover the full vocabulary of the diagnostic dataset. Therefore, the analysis in terms of psycholinguistic features is not exhaustive. The results are shown in Appendix .

In general, most concepts in the property datasets refer to concrete concepts and tend to have high ratings for all three features. What can be observed across all ratings is that the property datasets for **fly**, **dangerous**, and **lay_eggs** contain concepts that score low for all three features compared to other property datasets. This is likely to be caused by polysemous terms that can, for instance, be bird names, but also have other, more abstract senses (e.g. *swift*, *ruff*). A difference in classes can be observed for the property dataset **cold** for both concreteness and imageability. Negative concepts tend to score lower for the two features than positive concepts.

7.4.4 Relation Profiles

Finally, property datasets can be characterized in terms of the property-concept relations they contain. Each property could have its own 'relation profile'; for instance, it could be expected that color properties tend to be much more variable than part properties. To get insights into such tendencies, I calculate the proportion of positive examples annotated with a relation.

Table 7.12 shows an overview of relations over property datasets. For this analysis, only positive examples were considered, as the purpose of the relations is to characterize the relationship between properties and concepts they apply to. The top relations for each property are shown in bold (top proportion per relation and proportions closest to it). The analysis gives rise to several patterns:

Perceptual properties Overall, it can be observed that all perceptual properties tend to have a high proportion of variability relations. For three out of five color properties (**black**, **blue**, and **red**), it seems that the variability relations are, in fact, the only prominent relations. In other words, these three color properties do not tend to have strong, positive associations with concepts. Rather, they tend to apply to a subset of concept instances. For other perceptual properties, this pattern is less extreme: taste, temperature, material, and shape properties tend to be afford activities (`affording_activities`). For instance, 70% of positive examples in the **juicy** dataset are labeled as `affording_activity`, compared to 7% in the dataset for **black**.

		implied_category	typical_of_concept	typical_of_property	affording_activity	afforded_usual	afforded_unusual	variability_limited	variability_open	rare	unusual	impossible	creative
activities	fly	**0.89**	**0.91**	0.55	-	**0.91**	0.03	0.12	-	0.02	0.02	0	0
	lay_eggs	**0.97**	0.69	0.29	-	**0.96**	0.08	0.09	-	0	0	0	0
	roll	**0.62**	0.16	0.16	-	0.33	**0.56**	0.45	-	*0.23*	*0.21*	0.02	0.26
	swim	**0.76**	**0.58**	0.41	-	**0.64**	0.34	0.08	-	0.02	0.03	0	0
complex	dangerous (scalar)	**0.87**	**0.83**	0.68	0.70	-	-	0.19	0.51	0.09	0.12	0	0.02
	used_in_cooking	**0.92**	**0.92**	0.62	**0.92**	-	-	**0.99**	0.05	0.02	0.02	0	0
parts	wheels	**0.88**	0.78	0.44	**0.90**	-	-	0.12	0.09	0.02	0.03	0.02	0.02
	wings	**0.95**	**0.95**	0.65	**0.93**	-	-	0.04	0	0	0	0	0
perceptual-color	black	0.16	0.36	0.14	0.06	-	-	**0.70**	0.32	*0.26*	*0.25*	0	0.02
	blue	0.05	0.22	0.07	0.07	-	-	0.48	**0.62**	*0.30*	*0.28*	0	0
	green	0.46	0.62	0.38	0.36	-	-	**0.80**	0.23	0.12	0.11	0	0.01
	yellow	0.44	**0.77**	0.47	0.16	-	-	**0.79**	0.21	-	-	-	-
	red	0.14	0.36	0.15	0.11	-	-	**0.70**	0.55	*0.21*	*0.29*	0	0.12
perceptual-shape	round	**0.59**	0.35	0.16	0.32	-	-	**0.62**	0.53	0.11	0.11	0	0.09
	square	0.46	0.49	0.11	0.27	-	-	**0.83**	0.54	0.06	0.03	0	0
perceptual-temperature	warm (scalar)	0.47	0.47	0.32	0.50	-	-	0.02	**0.89**	0.1	0.1	0	0.02
	cold (scalar)	0.27	0.31	0.23	0.27	-	-	0.00	**0.79**	0.09	0.14	0	0
	hot (scalar)	0.51	0.63	0.41	0.63	-	-	0.11	**0.87**	0.09	0.09	0	0.01
perceptual-taste	sweet	0.58	0.63	0.33	0.61	-	-	**0.93**	0.11	0.08	0.12	0	0.01
	juicy	0.66	0.63	0.45	0.70	-	-	**0.93**	0.08	0.01	0.01	0	0
perceptual-material	made_of_wood	0.55	0.58	0.32	0.42	-	-	**0.92**	0.23	0.13	0.1	0	0

Table 7.12: Distribution of relations over positive examples of properties.

Part properties A different pattern can be observed for part properties: Both properties of this property type show low proportions of variability relations. Rather, the properties seem to apply to all instances of concepts. For some concepts, the property may even apply to a larger semantic category they are part of, as can be observed by high proportions for the relation `implied_category` (e.g. 95% for the property **wings**). Furthermore, part properties tend to afford activities. This is highly plausible, as wings and wheels tend to fulfil functional purposes.

Complex properties Complex properties are properties that arise from a combination of factors and depend on interpretation (e.g. multiple factors in combination lead to the fact that *tigers* are interpreted as dangerous animals.) Complex properties also show high proportions of relations that apply to most or all instances of a concept and tend to be relevant for functions or activities. In contrast to part properties, they show higher proportions of variability.

Activity properties The four activity properties all share a relatively high proportion of the relation `implied_category`. This can be seen as an indication that our conceptual systems are organized around actions, as argued by Borghi and Caramelli (2003). The two relations **fly** and **lay_eggs** have high proportions of the relation `afforded_usual` and low proportions of `afforded_unusual`. In contrast, **roll** is characterized by `afforded_unusual`. **Swim** follows a similar pattern as **fly** and **lay_eggs**. This is plausible, as the four properties differ with respect to how strongly they are tied to specific taxonomic categories: Laying eggs and being able to fly is limited to specific animals, whereas many things can roll, even if this is not what they normally do or are used for. This is also reflected by a high proportion of variability for the property **roll**. The ability to swim is also tied to taxonomic categories, but less strongly than flying or laying eggs. Mammals tend to be able to swim, but do not necessarily usually engage in this activity (e.g. cats).

	examples
blue	ring flame piano hose taxi frog wand currant glass night pot iceberg scorpion recorder football moss feather sandpaper
red	vinegar ring ginger rooster currant mangifera couch clarinet wand hair jeep aubergine mango glass miner apricot football melon onion chameleon nectarine fern oven syrup lewisia squirrel
black	pelican ring cherry leopard pea hornet owl sheep pipefish lizard nightmare crocodile rhino potoroo glass football eye opossum acaridae weasel coatis snake raccoon paridae zebra pig
roll	propeller footrest tub saw pipe tappet glass lathe washer lever bottle car plastic bucket bearing nut dowel

Table 7.13: Positive examples with negative relations.

Finally, the table indicates that for several properties, positive examples have been annotated with negative relations, in particular `rare` and `unusual`. The properties **black**, **red**, **blue**, and **roll** show comparatively high proportions for these relations. In practice, these relations can only apply to concept-property pairs that do not have relations in the ALL-SOME category, as such examples are treated as unreliable and not classified as either positive or negative examples. It is plausible that some concepts have been annotated with one of the variability relations as well as one of the negative relations. This can be the case for property-concept combinations that are not well known or allow for different interpretations. Table 7.13 shows all positive examples for which negative relations have been annotated for the four properties **black**, **red**, **blue**, and **roll**. Overall, the examples seem to confirm the hypothesis; several examples are valid positive examples, but do require more specific knowledge or at least familiarity (e.g. **blue**-*flame*, **blue**-*frog*, **black**-*pelican*). Other examples illustrate instances in which the property is vague given the property-concept combination and therefore difficult to interpret (**red**-*squirrel*, **black**-*zebra*, **black**-*cherry*). For the property **roll**, this vagueness is particularly prominent.

7.5 Property Profiles

In this section, I provide an assessment of each property dataset based on the factors described in the previous sections. In particular, I consider the effects of combinations of different risk factors (Section 7.5.1) and the potential explanatory value of each property dataset (Section 7.5.2).

7.5.1 Overview of Risks

Based on the factors considered above, it is possible to draw conclusions about the degree to which different property-datasets can yield insights about property-representation by distributional models. To summarize the risks for diagnostic experiments, I classify the different factors observed above into the following four categories:

Difficulty Various factors are likely to pose a particular challenge to diagnostic classification. Among them are a high number of low frequency words in the property dataset, a high degree of ambiguity in the entire set or one of the classes, and a high class imbalance. Rather than posing a risk for misleading positive results, difficulty may offer an explanation for negative results.

Noise Noise in the dataset can be caused by low quality annotations (indicated by low inter-annotator agreement) and the direct identification of vague or false-positive examples.

Separability A major risk of diagnostic classification experiments is that the positive class can be distinguished from the negative class by relying on aspects other than the target information. In the case of semantic properties and concepts, such a situation can occur if all positive examples happen to form a coherent semantic category (e.g. RED FRUITS) from

property	total score	difficult	noise	correlation	separability
fly	0				
green	0				
sweet	0				
juicy	0				
yellow	0				
hot	0				
black	1		x		
blue	1		x		
warm	1	x			
cold	1	x			
made_of_wood	1	x			
swim	1	x			
wheels	2	x			x
dangerous	2	x			x
red	2		xx		
square	2	x		x	
wings	2				xx
round	3	x	x		x
used_in_cooking	4	x		x	xx
roll	4	x	xx	x	
lay_eggs	4	xx		x	x

Table 7.14: Overview of risk factors per property dataset: Individual risk factors within a risk category are indicated by 'x'. The total score is the sum of identified risk factors.

which all all negative examples can easily be distinguished. This risk was assessed by means of a clustering analysis and by means of measuring cosine distances from the centroid vector over the positive class.

Correlation Accidental correlations with a class can be caused by many linguistic and distributional factors and are difficult to control. One factor that can be explored is frequency. Stark differences in word frequencies between the positive and negative class are considered a risk of accidental correlation.

Table 7.14 provides an overview of all four risk factors for each property dataset. The number of factors that apply within a risk category are indicated by crosses ('x'). For example, if a property dataset has low inter-annotator agreement and potentially noisy or vague false positives, this is indicated by two crosses in the column for 'noise' (as is the case for **red**). The total score indicates how many risk factors apply to a property dataset. It can be observed that for six properties, no risk factors have been identified. For another six properties, only one risk factor applies (either noise or difficulty). For nine property datasets, 2 or more factors apply.

It should be noted that not all risk factors carry the same weight. For instance, a high degree of separability is not necessarily an indication of an accidental correlation; rather, it may be caused by a particular strong encoding of the target property (**wings**, **wheels**). The

combination of a potential correlation with frequency and a high degree of separability, does, however, pose a considerable risk (**lay_eggs**, **used_in_cooking**). Class imbalance by itself is not necessarily problematic either. In combination with other factors that make correct classification difficult (low frequency and high ambiguity) or noise (**round**, **roll**), it may, however, pose the risk that information cannot be identified by diagnostic classifiers even if it is encoded in the embedding representations.

7.5.2 Explanatory Power

Based on the relation profile of each property and its risks in a diagnostic setup, it is possible to assess the suitability of individual property datasets to provide insights in diagnostic experiments. In an ideal scenario, individual property datasets would have a single, salient relation that applies to most positive examples. To test for possible interactions, the results of the property dataset could be compared to another property dataset that allows for controlling possible interactions.

Consider the following example: Implied information is expected to be not made explicit in corpus data and should therefore not be encoded in distributional representations. The ideal scenario to test whether this is the case would be a dataset in which the majority of positive examples is only annotated with the relation `implied_category`, which is not the case for any of the property datasets. The dataset that comes closest to this ideal is the dataset for the property **swim** (see Table 7.12): 76% of positive examples are annotated with the relation `implied_category`. However, large proportions of these examples are likely to be annotated with other relations that would interfere with the analysis. For example, high proportions can also be found for the relations `afforded_usual` (64%), and `typical_of_concept` (58%). A single property-concept pair can be annotated with multiple relations which causes overlap between pairs assigned to different relations. Unfortunately, no other property dataset allows for a comparison to control for this; the dataset for **fly** comes close, but has a higher proportion of examples for all three relations. Furthermore, drawing conclusions from comparing two property datasets does not exclude the possibility that the reason for the performance difference lies in the properties themselves. Ideally, properties of the same property type with different relation profiles should be compared. Unfortunately, it is not possible to find configurations in the dataset that allow for isolating the effect individual semantic relations.

To illustrate the complexity of potential interactions between relations, the most frequent relation configuration for each semantic property dataset is shown in Table 7.15. In addition to the most common relation configuration, the table shows the proportion of positive examples that share the configuration. For most relations, the proportion of the most common configuration is comparatively low (e.g. 9% for **roll**, below 50% for 18 out of 21 properties). This indicates considerable diversity in the configurations of relations. Furthermore, for most properties, many different relations are part of the configuration, making it almost impossible to isolate the effect of individual relations. Thus, at this point, the dataset is not suitable to give fine-grained insights into specific factors that may impact the representation of semantic properties in embedding vectors based on diagnostic experiments.

	top-config	proportion
roll	afforded_usual implied_category typical_of_concept typical_of_property	0.09
red	variability_open	0.10
round	variability_open	0.10
square	variability_limited variability_open	0.12
made_of_wood	affording_activity implied_category typical_of_concept typical_of_property variability_limited	0.17
black	variability_limited	0.18
yellow	implied_category typical_of_concept typical_of_property variability_limited	0.19
hot	affording_activity implied_category typical_of_concept typical_of_property variability_open	0.24
green	affording_activity implied_category typical_of_concept typical_of_property variability_limited	0.24
sweet	affording_activity implied_category typical_of_concept typical_of_property variability_limited	0.25
warm	variability_open	0.31
dangerous	affording_activity implied_category typical_of_concept typical_of_property	0.31
blue	variability_open	0.32
swim	afforded_usual implied_category typical_of_concept typical_of_property	0.39
lay_eggs	afforded_usual implied_category typical_of_concept	0.39
wheels	affording_activity implied_category typical_of_concept typical_of_property	0.41
juicy	affording_activity implied_category typical_of_concept typical_of_property variability_limited	0.41
fly	afforded_usual implied_category typical_of_concept typical_of_property	0.46
cold	variability_open	0.53
used_in_cooking	affording_activity implied_category typical_of_concept typical_of_property variability_limited	0.57
wings	affording_activity implied_category typical_of_concept typical_of_property	0.63

Table 7.15: Most frequent relation-configuration for each property dataset and the proportion of positive examples it is shared by.

7.6 Discussion and Conclusion

This chapter has provided a characterization of the diagnostic dataset. Particular focus has been placed on assessing to what degree it is a suitable instrument for diagnostic classification experiments and whether it can provide insights into underlying tendencies that govern whether property information is encoded in distributional data.

The assessment of the suitability for diagnostic classification has focused on several aspects: difficulty (balance between the classes, word frequency, degree of ambiguity), chance of noise (agreement, vague or false-positive examples), potential accidental correlation with frequency, and the distribution of positive and negative examples in the distributional space.

Each of these factors can pose a potential risk for a misleading outcome. For six property datasets, no risks could be identified. Six further property datasets show one risk factor either relating to potential noise or difficulty. The remaining nine property datasets show several risk factors involving the chance of accidental correlation and a high degree of class separability. The combination of the two factors applies to two datasets and warrants particular caution when interpreting the results of diagnostic experiments.

Despite these limitations for some of the property datasets, the diagnostic dataset can still be considered an improvement over approaches that purely rely on extracting examples from feature norm datasets (e.g. Fagarasan et al., 2015), as it contains verified negative examples. The dataset can be seen as a complementary resource to the Quantified McRae norms (Herbelot and Vecchi, 2016); it also presents information about subsets of concept instances a property applies to. The current dataset contains fewer properties than other feature-norm datasets. Instead, it contains more substantial sets of positive and negative examples per property. Furthermore, to the best of my knowledge, the analysis of suitability and risks for diagnostic classification is unique to the diagnostic dataset presented in this thesis.

The analysis of fine-grained property-concept relations in the datasets indicates complex interactions between relations that do not allow for controlled analyses. The relations do, however, provide rich information into property-concept combinations and can be considered a resource of fine-grained aspects of common sense knowledge. In future research, it could be considered to extend the dataset in such a way that individual hypotheses can be tested in diagnostic experiments.

7.7 Summary

The main purpose of this chapter was to present a characterization of the diagnostic dataset and analyze it with respect to its suitability and explanatory power in diagnostic experiments. The chapter presented a global analysis of the core aspects of the dataset (properties, concepts, and their relations), as well as a detailed analysis of the 21 property datasets. The different property-datasets constitute suitable diagnostic tools to varying degrees. In contrast to other resources, the dataset presented in this thesis contains substantial sets of verified positive and negative examples as well as fine-grained information about its particular risks and shortcomings on the level of individual properties.

The analysis also indicated that at this stage, the property-concept relations are not yet able to indicate information about the potential underlying factors that determine whether property-information is encoded in distributional data. Nevertheless, the relations provide rich information about specific property-concept pairs and thus constitute a resource of fine-grained common sense knowledge.

Part IV

Experiments

Part V of this thesis presents experimental work based on the diagnostic dataset. First and foremost, the dataset was designed for diagnostic experiments on context-free embedding representations. Chapter 8 presents two experimental setups specifically designed to tackle the problems of diagnostic classification by means of exploiting the strengths of the diagnostic dataset using context-free word vectors. The results indicate that context free embedding models are unlikely to capture property-specific information.

To complement these experimental results, I conduct a corpus analysis (presented in Chapter 9) of two corpora underlying the context free models to explore what type of linguistic property-evidence they contain and whether this evidence is likely to be represented by the embedding vectors and identified by diagnostic classifiers. The insights obtained in this analysis confirm that property-specific evidence does not provide a strong signal that is likely to end up in embedding representations. In contrast, fine-grained semantic category information seems to be more salient. I also explore to what degree the linguistic evidence found in the corpora is in line with the hypotheses presented in Chapter 3.

The final chapter (Chapter 10) of this part presents initial steps towards analyzing semantic property knowledge captured by contextualized language models. I use the diagnostic dataset to design two tasks that require semantic property knowledge. Rather than using diagnostic classification, I chose to rely on model behavior, namely masked token prediction for pre-trained language models and a common-sense reasoning task based on Winograd sentences for fine-tuned language models. While the pre-trained language models do seem to capture at least some semantic properties, this knowledge is not reflected by the fine-tuned models. Rather, they seem to be relying on different types of discourse structures when making decisions.

8. Diagnostic Classification of Context-free Models

8.1 Introduction

This chapter presents two probing studies that aim to detect to what degree semantic properties are encoded in context-free embedding representations. The core idea behind both probing experiments is the following: If information about a semantic property is encoded in the embedding representations, it should be learnable by a simple, binary classifier trained on positive and negative examples of the property. For instance, if the property **fly** is encoded in the embedding representations, a binary classifier trained on positive examples of the property (e.g. *pigeon*, *eagle*, *bee*, *wasp*, *airplane*) and negative examples of the property (e.g. *ostrich*, *table*, *car*, *spider*, *boat*) should be able to distinguish held out positive examples (e.g. *seagull*,*butterfly*, *helicopter*) from held out negative examples (e.g. *penguin*, *caterpillar*, *boat*).

Probing has a major methodological challenge: Probing results by themselves cannot necessarily provide conclusive results about whether a piece of information is encoded as it is difficult to verify that the distinguishing features identified by the classifier are indeed linked to the target information. At the core of the problem lies the possibility of accidental correlations in the data. This possibility complicates the interpretation of probing performance:

Correlation In many cases, semantic properties correlate with semantic categories. If the majority of positive examples of a property (e.g. **fly**) is taken from the same category (e.g. BIRD) and the majority of negative examples is taken from radically different categories (e.g. FURNITURE, CLOTHING), it is impossible to determine whether the classifier learned to identify information about the property or whether it learned to recognize the semantic category represented in the positive class.

Interpretation It can be expected that probing classifiers will hardly ever reach perfect performance, as information is unlikely to be encoded sufficiently for every single instance in the training and test data. However, imperfect performance might also be caused by other factors: The classifier may have learned to identify other aspects of semantic information that happen to correlate with the positive and negative class, such as semantic categories or other, accidental correlating aspects. High, but not perfect performance by itself can thus not be a guarantee that the classifier could, indeed, identify the target property.

The problem can be addressed from two perspectives: It is possible to control the distribution of positive and negative examples to avoid obvious correlations. At the same time, it is possible that accidental correlations remain present. Therefore, a complementary approach is to compare classifier performance to baselines that do not require the identification

of the target property to yield reasonable performance. If a probing classifier can outperform such a baseline, this is a good indication that it could go beyond general semantic similarity (e.g. between the members of a category) and managed to identify property-specific information in the embeddings.

In this chapter, I present two studies that aim to address these challenges in their methodological setup. The first study (Section 8.2) uses a pilot version of the dataset and compares probing classifiers to a classification approach based on similarity to an approximated property embedding. This comparison allows to gain insights into whether a classifier could access individual vector dimensions and thus go beyond general similarity. The results provide initial indications that visual-perceptual properties are most likely not encoded in embeddings, while properties relevant for how entities interact with the world (functions, actions) may well be encoded.

The second study (Section 8.3) uses the full version of the diagnostic dataset and can thus exploit the controlled distribution of positive and negative examples: The positive examples are taken from a variety of semantic categories and negative examples have high similarity to positive examples. A classifier thus has to go beyond general similarity in order to perform well. In addition, the study employs a control and ceiling task. The control task aims to determine to what extent the probing classifiers could indeed learn to abstract over property information (rather than classify examples based on general, not property-specific similarity to training set examples). The ceiling task indicates to what extent information can be learned given the size and class distribution of a property set. The results indicate that property information is not encoded for color properties. Other properties yield better results on the probing task when compared to the control task, but the error analysis indicates that the classifiers learn fine-grained semantic categories, rather that property-specific information.

The first study (presented in Section 8.2) is based on the following publication:

> Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286

8.2 Study 1: Probing vs. Nearest Neighbors

The study presented in this section uses a pilot version of the diagnostic dataset and offers a first methodological set-up to derive meaningful results from probing experiments on embedding representations. The main focus lies on comparing probing classifiers to a classification approach based on cosine similarity to an approximated property representation. The probing classifiers should be able to access individual vector dimensions encoding specific aspects of semantic information (i.e. subspaces of the embedding vectors). The cosine-similarity approach cannot access subspaces and thus remains limited to comparing embedding in terms of general similarity.

If property-specific information is encoded in the embedding representations, the probing classifiers should be able to detect the relevant subspaces, even if the representations of the example concepts come from radically different semantic categories and are spread over

the entire semantic space (e.g *lipstick*, *blood* and *strawberry* should all carry information of the property **red**). In contrast, the nearest neighbor approach will only perform well on properties whose positive examples are taken from a relatively coherent semantic category (e.g. examples that share a taxonomic category such as BIRD: *pigeon*, *sparrow*, *eagle*). It will not perform well on properties that cut across a diverse set of semantic categories.

The comparison of the two approaches can be interpreted as follows: If both the nearest neighbor approach and the probing classifiers yield low performance for a property, it can be concluded that property information is not present in the embeddings. In contrast, if the nearest neighbor approach is outperformed by the probing classifiers, it can be concluded that it could detect information that goes beyond general semantic similarity. The third possibility is that both approaches perform equally highly. This outcome cannot provide indications about whether a property is encoded in the embeddings.

The focus of the study lies on exploring the representation of individual semantic properties. In addition, we[1] also experiment with property-encoding for words involved in regular polysemy. We test specific hypotheses about what type of property-information we expect to be encoded in embeddings. While we find mixed evidence for several aspects, the results seem to provide initial indications that visual-perceptual properties are most likely not encoded in embeddings. Properties that relate to the way entities interact with the world (expressed as actions or functions) do seem to be identified by the probing classifiers.

The subsequent sections are taken from the original publication. Minor adaptations have been made to integrate the text in the larger framework of the thesis. The remainder of this section is structured as follows: The details of the method are outlined in Section 8.2.1. Section 8.2.2 presents our experiments and results. We finish with a critical discussion and overview of future work in Section 8.2.3.

8.2.1 Method

The core of our evaluation consists of testing whether nearest neighbors and classifiers are capable of identifying which embeddings encode a given semantic property. We first describe the dataset and then present the procedure we apply. We complete this section with our hypotheses about the outcome of our evaluation.

Extended CSLB Data

The dataset used for this study is based on the CSLB norms and constitutes a pilot version of the diagnostic dataset presented in Part III. The pilot dataset consists of examples from the CSLB set and has been extended via an embedding model. In a second set, we used crowd-sourcing and manual verification to select appropriate negative examples. In this section, we outline the steps taken to construct the dataset and highlight its most important characteristics. The details of the dataset construction are described in Chapter 5. A record of the decisions taken in the annotation and verification process can be found on Github.[2]

[1] The work presented in this Section has been conducted together with Antske Fokkens. The experimental set-up was developed together and implemented by me. The publication was co-written by both authors.

[2] All annotations, guiding principles as well as notes about resolving discussions can be found at `https://cltl.github.io/semantic_space_navigation`.

Classification approaches

We use the pretrained Word2vec model based on the Google News corpus.[3] The underlying architecture is a skip-gram with negative sampling model (Mikolov et al., 2013b), which learns word vectors by predicting the context given a word.

The overall goal is to investigate whether word vectors capture specific semantic properties or not. We start from the assumption that classifiers can learn properties that are represented in the embedding in a binary classification task. We apply supervised classification to see whether a logistic regression classifier or a neural network are capable of distinguishing embeddings of words that have a specific semantic property from those which do not. Specifically, we use embedding vectors corresponding to words associated or not associated with a semantic target-property (i.e. positive and negative examples) as input for a binary classifier and test whether the classifier can learn to distinguish embeddings of words that have the property from those who do not. However, word embeddings also capture semantic similarity. If a property is shared by similar entities (e.g. most animals with a beak are birds), the classifiers may perform well because of this similarity rather than identifying the actual property. We therefore compare the performance of classifiers to the performance of an approach based on full vector similarity. If only the classifiers score well, this provides an indication that the embedding captures the property. If both methods perform poorly this could mean that the property is not captured.[4]

Supervised classification As the datasets are limited in size, we evaluate our classifiers by applying a leave-one-out approach. We employ two different supervised classifiers, which we expect to differ in performance. As a 'vanilla' approach, we use a logistic regression classifier with default settings as implemented in SKlearn (Pedregosa et al., 2011). This type of classifier is also used by Drozd et al. (2016) to detect words of similar categories in an improved analogy model.

In addition, we use a basic neural network. Meaningful properties may not always be encoded in individual patterns, but rather arise from a combination of activated dimensions. This is not captured well by a logistic regression model, as it can only react to individual dimensions. In contrast, the neural network can learn from patterns of dimensions. We use a simple multi-layer perceptron (as implemented in SKlearn[5]) with a single hidden layer. We calculate the number of nodes in the hidden layer as follows: (number of input dimensions + number of output dimensions) * 1/3. The pretrained Google News vectors have 300 dimensions, resulting in a hidden layer of 100 nodes. We use the recommended settings for small datasets. No parameter tuning was conducted so far due to the limited size of the datasets and the use of a leave-one-out evaluation strategy. We present the runs of several models, as the neural network can react to the order in which the examples are presented as well as the randomly assigned vectors for initialization. While the performance of the

[3] `https://code.google.com/archive/p/word2vec/`

[4] Given the size and balance of our dataset as well as the lack of fine-tuning, we remain careful not to draw firm conclusions at this point.

[5] `http://scikit-learn.org/stable/index.html`

model could be optimized further by experimenting with the settings, we find that the setup presented here already outperforms the logistic regression classifier in many cases.

Full vector similarity To show that supervised classification can go beyond full vector comparison in terms of cosine similarity, we compare the performance of the classifiers to an n-nearest neighbors approach. We calculate the centroid vector of all positive examples in the training set. The training set consists of all positive examples in the leave-one-out split except for the one we are testing on. We then consider its n-nearest neighbors measured by their cosine distance to the centroid as positive examples. We vary n between 100 and 1,000 in steps of 100. We report the performance of the optimal number of neighbors for each property (which varies per property).

Variety approximation The performance of the approaches outlined above depends on to the variety of words associated with a property. We approximate this variety by calculating the average cosine similarity of words associated with a property to one-another. This is done by averaging over the cosine similarities between all possible pairs of words. A high average cosine similarity means that the words associated with a concepts tend to be close to each other in the space, which should mostly apply to words associated with taxonomic categories. In contrast, a low average cosine means a high diversity, which should largely apply to general descriptions.

Specific hypotheses

We select a number of properties for closer investigation based on the clean and extended dataset described in Section 8.2.1. We first formulated the hypotheses independently, before discussing and specifying them.[6] Table 8.1 summarizes the agreed upon expectations. The hypotheses can be categorized in the following way:

Sparse Textual Evidence We select properties of which we expect that textual evidence is too sparse to be represented by distributional vectors. The properties **is_black**, **is_yellow**, **is_red** and **made_of_wood** have little impact on the way most entities belonging to that class interact with the world. We expect that the only textual evidence indicating them are individual words denoting the properties themselves (e.g. **red**, **black**, **wooden**)[7] and it is unclear how often they are mentioned explicitly. It may, however, be the case that certain subcategories in the datasets are learned regardless of this sparsity, because they happen to coincide with more relevant taxonomic categories such as *red fruits*.

Fine-grained Distinctions in Larger Categories We expect that a supervised classifier may be able to make more fine-grained distinctions between examples of the same category when these differences are relevant for the way they interact with the world. We select

[6] Details can be found on the Github repository.

[7] In the case of **made_of_wood**, the evidence may be a bit broader, as it might be indicated by different types of wood occurring in the context of furniture.

Property	learnable property
is_an_animal	yes
is_food	yes
is_dangerous	yes
does_kill	yes
is_used_in_cooking	yes
has_wheels	possibly
is_found_in_seas	possibly
is_black	no
is_red	no
is_yellow	no
made_of_wood	no

Table 8.1: Hypotheses about whether selected semantic properties can be learned by a supervised classifier.

two properties that introduce crucial distinctions in larger categories: **has_wheels** and **is_found_in_seas**. The former applies to a sub-group of vehicles and may be apparent in certain behaviors and contexts only applying to these vehicles (rolling, street, etc). The latter applies to animals, plants and other entities found in water, but it is unclear whether textual evidence is enough to distinguish between seawater and fresh water.

Mixed Groups We expect that a supervised machine learning approach can find positive examples of a property that are not part of the most common class in the training set. For instance, the majority of positive examples for **is_dangerous** and **does_kill** refer to weapons or dangerous animals. We expect the classifier to (1) find positive examples from less well represented groups and (2) be able to distinguish between positive and negative examples of a well-represented category (e.g. **rhino** v.s. *hippo* for killing). For the property **is_used_in_cooking**, the example words refer to food items as well as utensils. We expect that classifiers can distinguish between cooking-related utensils and other tools.

Polysemy We expect that machine learning can recognize vector dimensions indicating properties applying to different senses of a word, whereas the nearest-neighbors approach simply assigns the word to its dominant class. We used the systematic sense shift between animal and food senses (a case of metonymy) to test this hypothesis. For instance, we expect that word vectors that can be used to describe animals as well as food (e.g. *chicken*, *rabbit* or *turkey*) record evidence of both contexts, but end up closer to one of the categories. A supervised machine learning approach should be able to find the relevant dimensions regardless of the cosine similarity to one of the groups and classify the word correctly. We test this by training on a set of monosemous words (animals and food items) and test on a set of polysemous and monosemous examples.

8.2.2 Experimental Setup and Results

Concept Diversity vs. Performance

We first investigate the relation between performance and diversity of concepts associated with a property on the full, noisy dataset using a leave-one-out approach. Table 8.2 shows a selection of the f1-scores achieved on properties in the CSLB dataset in relation to the average cosine similarity of the associated words. A high average cosine similarity means that the concepts overall have similar vector representations and can thus be seen as having a low diversity. The results of the Spearman Rank correlation clearly indicate that scores achieved by nearest neighbors correlate more strongly with the average cosine than the two supervised classification approaches. In addition, this the mean cosine similarity of positive examples to the centroid representations ('cos') indicates the diversity of examples associated with a property. For instance, the perceptual properties **is_heavy** and **is_thin** and the encyclopedic property **is_strong** have low mean cosine similarities (0.15, 0.16, 0.15), indicating that positive examples of these properties are scatter over the embedding space. This is plausible, as these properties can apply to a wide variety of concepts from many different semantic categories. In contrast, the property **does_make_music** has a high mean cosine similarity (0.55), indicating that its positive examples are all located close to one another in the embedding space. This is also plausible, as the property mainly applies to musical instruments, which form a coherent semantic category.

feature	cos	f1-neigh	f1-lr	f1-net	type
is_heavy	0.15	0.15	0.17	0.21	op
is_strong	0.15	0.13	0.13	0.34	e
is_thin	0.16	0	0.05	0.1	vp
is_hard	0.16	0.15	0.08	0.26	op
is_expensive	0.16	0	0.28	0.37	e
made_of_wood	0.17	0.14	0.62	0.62	vp
...	...	...	...	...	
is_black	0.2	0.29	0.23	0.24	vp
is_electric	0.21	0.48	0.5	0.69	vp
is_dangerous	0.21	0.53	0.57	0.59	e
is_colourful	0.21	0.14	0.25	0.32	vp
is_brown	0.21	0.13	0.22	0.33	vp
has_a_handle _handles	0.22	0.44	0.57	0.58	p
has_a_seat _seats	0.22	0.43	0.3	0.48	p
does_smell _is_smelly	0.22	0.08	0.15	0.37	op
made_of_glass	0.22	0.29	0	0.28	vp
has_a_point	0.23	0.38	0.23	0.47	p
does_protect	0.24	0.38	0.26	0.37	f
is_yellow	0.24	0.22	0	0.23	vp
is_soft	0.24	0.12	0	0.16	op
is_red	0.25	0.34	0.13	0.27	vp
is_fast	0.25	0.3	0.31	0.48	vp
is_tall	0.25	0.43	0.57	0.65	vp

is_a_tool	0.26	0.5	0.51	0.47	t
does_kill	0.28	0.64	0.48	0.62	f
...	...	...	...	...	
is_a_weapon	0.3	0.74	0.56	0.63	t
is_green	0.31	0.45	0.45	0.45	vp
has_a_ blade_blades	0.32	0.68	0.65	0.74	p
is_worn	0.32	0.47	0.86	0.9	f
has_wheels	0.32	0.82	0.83	0.87	p
is_found _in_kitchens	0.33	0.56	0.73	0.76	e
does_fly	0.33	0.57	0.76	0.76	f
has_a_tail	0.33	0.53	0.68	0.69	p
is_an_animal	0.33	0.64	0.76	0.78	t
is_eaten_edible	0.33	0.37	0.88	0.85	f
has_four_legs	0.34	0.67	0.66	0.66	p
is_a_vehicle	0.34	0.76	0.69	0.79	t
does_eat	0.34	0.68	0.71	0.68	f
...	...	...	...	...	
has_a_beak	0.37	0.63	0.83	0.87	p
made_of_cotton	0.37	0.68	0.56	0.64	vp
has_roots	0.37	0.3	0.65	0.72	p
is_a_mammal	0.37	0.69	0.85	0.86	t
does_grow	0.37	0.52	0.81	0.81	e
is_a_plant	0.37	0.43	0.63	0.64	t
has_leaves	0.37	0.41	0.71	0.78	p
...	...	...	...	...	
has_pips_seeds	0.47	0.5	0.08	0.46	p
is_juicy	0.5	0.71	0.48	0.56	op
is_a_vegetable	0.52	0.78	0.75	0.81	t
is_played _does_play	0.53	0.9	0.98	0.98	f
does_make_music	0.55	0.89	0.95	0.92	f
spearman-r		0.72	0.52	0.59	

Table 8.2: Performance of different approaches in relation to the average cosine similarity of words associated with a property (cos). The last row shows the Spearman Rank correlation between f1-scores and average cosine similarity. Property types are listed under type (p = part, vp = visual-perceptual, op = other-perceptual, e = encyclopaedic, f = functional, t = taxonomic).

Outcome of Specific Hypotheses

We carry out further experiments on a small extended and clean subset, consisting of carefully selected negative examples from the CSLB dataset and crowd annotations validated by the authors. The distribution of positive and negative examples per property is shown in Table 8.3. For some properties, the sets derived from the CSLB norms alone have an imbalanced distribution of negative examples over semantic categories, as they were selected by means of logical exclusion (e.g. concepts listed under **has_wheels** have been selected as negative

Property	pos	neg
full_does_kill	101	69
crowd_does_kill	67	49
full_has_wheels	79	349
full_is_black	42	89
full_is_dangerous	177	104
crowd_is_dangerous	131	84
full_is_found_in_seas	83	72
crowd_is_found_in_seas	47	28
full_is_red	29	80
full_is_used_in_cooking	142	61
full_is_yellow	24	68
full_made_of_wood	87	282
full_is_an_animal_test	37	20
full_is_an_animal_train	166	77
full_is_food_test	37	20
full_is_food_train	97	146

Table 8.3: Class distribution in dataset consisting of the clean datasets derived from the CSLB set and the additional crowd judgments (marked **full_**). For some properties, we included the dataset consisting of crowd-judgments only, as it is more balanced across semantic categories than the full set (marked **crowd_**). For all properties, a leave-one-out approach was applied to evaluation except for **is_animal** and **is_food**.

examples of **is_food**). Therefore, we add the more balanced but smaller datasets created by crowd-judgments only where enough judgments have been collected. We created additional sets for words part of the food-animal polysemy to test whether supervised classifiers can successfully predict semantic properties of various senses of polysemous words. In the following sections, we will outline the most striking results. Most results confirm, but some contradict our initial hypotheses.

Table 8.4 shows the f1-scores on the full clean datasets. As hypothesized, the color properties **is_yellow** and **is_red** perform low in all approaches, with slightly better results yielded by supervised learning.

The properties involved in functions and activities or with high impact on the interaction of entities with the world all perform highly in the classification approaches. For **does_kill**, **is_dangerous** and **is_used_in_cooking**, there is a large difference between the best nearest neighbors approach and the best classification approach (between 60 and 19 points), indicating that the classification approaches are able to infer more information from individual dimensions than is provided by full vector similarity. The property **is_dangerous** has, as can be expected, a particularly high diversity of associated words (comparable to the colors). **Has_wheels** and **is_found_in_seas** can be expected to have high correlations with other taxonomic categories (fish and water animals, vehicles), which is reflected in the lower diversity and comparatively high nearest neighbor performance.

Cases contradicting our expectations are the visual properties **is_black** and **made_of_wood**. Both have comparatively high classification performance with a big difference to the nearest neighbor results. Most likely, this is due to a category bias in the negative examples.

For instance, a large proportion of the negative examples for **is_made_of_wood** consist of animals and food. In the dataset for **is_black**, a large proportion of the positive examples consists of animals.[8] A classifier can perform highly by simply learning to distinguish these two categories from the rest.

The biases in semantic classes mentioned above partially result from the way we generated the negative examples from the original CSLB dataset. This means that a classifier may learn to distinguish two semantic categories rather than being able to find vector dimensions indicative of the target property. We therefore also present selected results on crowd-only datasets shown in Table 8.4, which do not have this bias. It can be observed that for all three properties,[9] the performance of the classification approaches drops marginally, whereas it rises for nearest neighbors.

We investigate the outcome on a number of individual examples to gain more insights into whether the subtle differences hypothesized in Section 8.2.1 hold. Since we only formulate a general hypothesis for Sparse Textual Evidence, we do not dive deeper into the results for that category here.

Fine-Grained Category Distinctions The full clean **has_wheels** dataset includes a number of instances for which the classifiers can make more fine-grained distinctions than nearest neighbors. As hypothesized, classifiers, in contrast to nearest-neighbors, can recognize that neither *sled* nor a *skidoo* have wheels, but a *unicycle* a *limousine*, a *train*, *carriage*, an *ambulance*, a *porsche* do. Another fine-grained distinction can be identified in the *is_found_in_seas* crowd-only set: *Sculpin* is correctly identified as a seawater fish by all classifiers but not by nearest-neighbors.

Mixed Groups Whereas nearest neighbors predominantly identify weapons as **is_dangerous** in the crowd-only set, the classifiers go beyond this category. The neural network approach correctly identifies that *imitation pistol*, *imitation handgun*, and *screwdriver* are negative examples of **is_dangerous**. Furthermore, no animals are labeled as dangerous based on proximity to the centroid, but the classifiers are able to distinguish between some dangerous and non-dangerous animals (e.g. *rhinoceros* is labeled positive, while *giraffe* and *zebra* are labeled as negative). All three classifiers recognize that *meth*, *cocaine* and *oxycodone* are considered dangerous substances, despite the fact that they are far away from the centroid of dangerous things. Of the only two disease-like concepts, *Hepatitis C* and *allergy*, the former is recognized by all classifiers and the latter only by logistic regression. The performance on the smaller, but also weapon-dominated **does_kill** crowd-only set is comparable, but the variety of atypical cases is lower. Among the only two disease-related items, *dengue* is identified by all classifiers and *dengue virus* only by the neural network.

In the crowd-only *is_found_in_seas* set, *seabird* and *gannet* are correctly labeled as positive, even though positive examples almost exclusively consist of fish or underwater-

[8]The property has a comparatively low mean cosine similarity between the centroid and the positive examples (0.19, as shown in Table 8.4), indicating high diversity of positive examples. In this particular case, it is possible that the low mean is caused by outliers rather than a generally diverse distribution of examples.

[9]We only included properties for which we had enough positive and negative examples in our set

property	av-cos	neigh	lr	net1	net2
full_is_yellow	0.23	0.19	0.47	0.64	0.64
full_is_used_in _cooking	0.37	0.29	0.98	0.98	0.98
full_is_black	0.19	0.35	0.75	0.77	0.77
full_is_red	0.23	0.36	0.51	0.54	0.52
full_is_dangerous	0.24	0.58	0.88	0.88	0.87
crowd_is_dangerous	0.26	0.61	0.86	0.86	0.86
full_has_wheels	0.38	0.90	0.96	0.96	0.95
full_is_found_in_seas	0.44	0.87	0.97	0.98	0.98
crowd_is_found _in_seas	0.50	0.87	0.94	0.96	0.96
full_does_kill	0.27	0.67	0.83	0.86	0.82
crowd_does_kill	0.30	0.70	0.82	0.84	0.80
full_made_of_wood	0.17	0.14	0.84	0.85	0.85
full_is_food_test	0.37	0.00	0.36	0.36	0.36
full_is_an _animal_test	0.37	0.52	0.88	0.88	0.88

Table 8.4: F1 scores achieved by logistic regression (lr) two runs of a neural net classifier (net1 and net2 and the n-best nearest neighbors evaluated with leave-one-out on the full datasets (marked as *full_* and the crow-only sets (marked as *crowd_*).

.

animals, whereas the negative examples encompass a vast variety of animals, including *bird* and some freshwater fish.

Polysemy For systematic sense shifts between food and animal senses of words (metonymy) (Table 8.4), we observe that when trained on pure animal and food words and tested on polysemous animal and food words, the classifiers perform highly with a large difference to nearest neighbors. For food versus pure animal words, the classifier performance is much lower. We expect the extremely low nearest neighbor performance to be due to the fact that the centroid is calculated over pure food items (without a single animal-related item, not even culinary meat terms such as *pork* or *beef*) which is far away from the animal-region in the space. Despite the classifiers outperforming nearest neighbors, the outcome does not confirm our original hypotheses. We expected that the classifiers could identify that edible animals have both animal properties and food properties, but upon inspection of the results, the classifiers only identified entities with a predominant animal sense correctly as animals and those with a predominant food sense correctly as food.

8.2.3 Discussion and Conclusions

The experiments presented in this approach have several limitations. First, our semantic datasets are still limited in size. Second, the implication method we applied to generate negative examples led to biases for some properties where most negative examples belong to a small set of (taxonomic) classes. Third, no parameter tuning has been carried out so far. Careful parameter tuning would ensure that the best possible classification approaches are chosen and that the obtained results truly exploit the informative power of the embeddings. Due to the limited size of the dataset and the leave-one-out approach to evaluation, this has

not been possible in this preliminary study. Fourth, the experiments presented here only concern a small subsection of semantic properties too limited to draw general conclusions. Despite these limitations, our results provide preliminary insights that lead us to conclude that the overall idea behind our methods works.

The main contribution of this study is that it introduces a new method aimed at investigating the kind of semantic information captured by word embedding vectors. We have taken the first steps towards constructing a dataset suitable for this investigation on the basis of an existing dataset of human-elicited semantic properties. We introduced a set of hypotheses concerning which semantic properties are captured by embeddings and presented exploratory experiments verifying them.

We show that classifiers, in particular neural networks, can identify which entities have a specific property in cases where this does not follow from general similarity or the overall semantic class the entity belongs to. This can be seen as a first indication that (some) semantic properties are encoded in individual (patterns of) vector dimensions, which can be identified.

The results on the extended datasets partly confirm that visual properties are not well represented by embeddings, while properties relating to function (e.g. cooking, having wheels) and interactions with other entities (e.g. being dangerous or killing) tend to be represented well. Some of these indications could be the result of the bias in our current dataset, but others have been confirmed on the smaller crowd-only sets for properties with enough available data (**is_dangerous** and **does_kill**). Further evidence is provided by the full dataset for **has_wheels** which encompasses a large group of vehicles to which the property does not apply. In addition, we support these indications by qualitative insights through examples of the kinds of distinctions made by the classifiers, but not the nearest neighbor approach. Results achieved for polysemous words and two visual properties currently do not confirm our hypotheses.

8.3 Study 2: Control and Ceiling Task

In this section, I present a study using the full diagnostic dataset. It addresses the methodological challenges of probing by means of two strategies: Firstly, we[10] introduce a control and ceiling task as tools to interpret classifier performance. Secondly, we use the full diagnostic dataset enables us to exploit the challenging example distribution in the property datasets.[11]

At its core, the study addresses the problem of how we can draw conclusions from a semantic probing task given the problems and limitations of diagnostic classification. As an illustration, consider the following scenario: To find out whether the property **fly** is encoded in embedding representations, we train a classifier on the following examples:

positive *airplane*, *rocket*, *eagle*, *pigeon*

negative *boat*, *emu*, *truck*, *penguin*

[10]The experiments in this study were designed in collaboration with Antske Fokkens and Piek Vossen. The experiments were implemented by me. The text in this section is partially based on an unpublished paper written in collaboration with Antske Fokkens and Piek Vossen.

[11]The code for the experiments and the datasplits used can be accessed via this repository: `https://github.com/PiaSommerauer/ControlledPropertyDiagnostics`

We test whether the classifier has learned to identify the property **fly** using the following test examples:

positive *helicopter*, *albatross*, *sparrow*

negative *rover*, *crane*, *ostrich*

The probing classifier provides the following output (wrong answers marked with an asterisk):

positive *helicopter*, *rover**, *sparrow*

negative *albatross**, *crane*, *ostrich*

The output allows for multiple interpretations: The first option is that the classifier learned to identify the target property **fly**, but the representation of *albatross* did not carry enough evidence of it. Likewise, the representation of *rover* happened to carry evidence of **fly**.

An alternative explanation of the outcome is that the classifier did not learn to identify evidence of the property **fly**. Instead, it found other similarities between train and test examples to arrive at its solution: The word *rover* is likely to appear in similar contexts as the positive training example *rocket*, as both words are related to space travel. The word *albatross* is likely to appear in similar contexts as the word *penguin*, as both birds are commonly found in Antarctica. The correctly classified examples can be explained by similar associations that do not necessarily require knowledge of the property **fly**: ostriches and emus share many properties and so do helicopters and planes, and trucks and cranes (in the sense of the lifting tool rather than the bird). In short, a classifier could have memorized examples in the training set and classified test set examples based on similarity to the memorized examples.

To determine whether a semantic property has been learned successfully by a diagnostic classifier, the second explanation for its output has to be ruled out. We employ two complementary strategies to distinguish between outcomes caused by memorization and outcomes caused by successful identification of the target property.

Example distribution. Classifying examples based on memorization is difficult if the examples in the positive and negative class are semantically diverse. Ideally, the positive examples only share the target property. The negative examples should only be connected by not having the target property. Given such a scenario, a classifier that only relies on memorization cannot achieve high performance.

Control task. A second strategy is to compare classifier performance to a strong baseline. A classifier that identifies the target property successfully should outperform a classifier that can only rely on memorization. A control task should contain examples that are connected by semantic similarity, but do not share the target property. For instance, a control dataset for the property **fly** could consist of the following examples:

positive *airplane*, *airport*, *albatross*, *penguin*, *seagull*, *ostrich*

negative *emu*, *crane*, *train*, *shuttle*, *satellite*, *orbit*

The classifier trained and tested on the dataset for **fly** should outperform the classifier trained on the control train and test set for **fly** if it managed to identify the target information. Both datasets should be equivalent with respect to size and class distribution.

The interpretation of low classifier performance also poses a problem: In practice, the property datasets struggle with an additional limitation that complicates the interpretation of the results: Various property datasets are limited in size and have an imbalanced class distribution. Thus, low classifier performance may either indicate that property-information is not learnable or that the target information is simply not learnable given the size and class distribution. We introduce a **ceiling task** to test whether information is learnable given a dataset with a particular size and example distribution.

Applying the control task and dataset analysis to the ceiling task shows that probing can detect information given a clear signal even with small imbalanced data. For most semantic properties, we only find weak or no evidence of information in the embeddings. Our analyses provide no evidence of visual property information in embeddings. Furthermore, they indicate that embeddings may represent taxonomic category information rather than property-specific information. These results are inline with previous findings (e.g. Rubinstein et al., 2015).

This section is structured as follows: We present methodological considerations with respect to the interpretation of results and the probing dataset for a semantic task in Section 8.3.1. We describe our experimental setup in Section 8.3.3 and present the results of our experiments in Section 8.3.4.

8.3.1 Methodological Considerations

In this section, we outline the methodological considerations underlying our experiments. As introduced in the previous section, major problem of diagnostic classification is the interpretation of classifier performance. If a classifier performs well (e.g. above a random or majority-class baseline), but does not achieve perfect results, this is not necessarily an indication that it identified the target information (Hewitt and Liang, 2019; Belinkov, 2021). It is impossible to distinguish between the following two underlying reasons:

1. **Imperfect representation**: The classifier learned to identify the property, but it is not represented in all instances. Errors are caused by instances in which the property is not represented.
2. **Memorization** or **Correlation**: The classifier did not learn to identify the target property, but performed above chance by classifying instances based on similarity to instances in the training data. The similarity can be caused by other semantic aspects that correlate with a class (e.g. a category). Hewitt and Liang (2019) emphasize that this risk of memorization increases with the complexity of the classifier. In this case, neither correct classifications nor errors indicate the presence or absence of the target information in the representation.

We test two complementary strategies to distinguish actual property learning from memorization: (1) Classifier selectivity using a control task and (2) a controlled distribution of property examples.

Classifier performance *below* baseline is not straight-forward to interpret either. It can indicate that property information is not present or that the size and distribution of the

diagnostic dataset are not sufficient for the classifier to learn it. We address this question by means of a ceiling task.

Classifier Selectivity

Hewitt and Liang (2019) propose control tasks against which the performance on the diagnostic target task can be compared. The goal of a control task is to determine how highly a classifier can perform if it can only rely on memorization of examples without having access to the information of interest (in our case semantic property information). Such a control task can be seen as baseline that can only be beaten by identifying the target information.

Hewitt and Liang create such a control task for probing part-of-speech (pos) information in contextualized language models by using the following strategy: They assign a random pos label to each word in the control set regardless of its context. The pos label is kept constant across all occurrences of a word in the training and test set. Since the pos assignment is random, the context in which tokens appear does not provide any indication of the correct pos label (e.g. the word *love* will always have the same label, regardless of its context and thus regardless of its actual pos label). Thus, the model representations used as input for the probe should not contain indicative information. This results in a situation in which a classifier can achieve reasonable performance by means of memorizing examples in the training set (i.e. occurrences of the same word) but not achieve perfect performance (unless there is 100% overlap between the training and test vocabularies).

If the classifier trained on the actual pos labels clearly outperforms the control classifier relying on memorization, this is a strong signal that it learned to identify pos information in the language model representations. This can be measured in terms of **selectivity**, which is defined as the performance difference between the target probe and the control probe.

Semantic control task. The notion of selectivity defined by Hewitt and Liang (2019) cannot be translated to our lexical semantic task directly. In our task, train and test splits never contain the same tokens. Memorization can still occur, but here it means that a classifier bases its predictions on general vector similarity to training examples instead of identifying the semantic target information.

A good control task for semantic property probing has to fulfill the following criteria: (1) Randomness: It should contain a set of randomly chosen examples assigned to the positive class of a property. Some of the examples will be positive instances of the property, but others will not. This means that a probe cannot rely on the target information anymore. (2) Similarity: The random positive examples should be connected by overall high vector similarity. This enables a probing classifier to exploit memorization while not having access to the target information. It can thus indicate what performance a classifier can achieve by means of classification purely based on similarity to memorized examples. This type of distribution constitutes a strong baseline. If the probe clearly performs higher on the real task than on the control task, we can interpret this as a strong indication that it could go beyond memorization.

We create control datasets by forming a chain of similar words by carrying out the following steps: (1) We randomly pick a word from the original set, which can be a positive or negative example. This word acts as our first seed of the positive control class. (2) We

look for nearest neighbor of the seed in an embedding model.[12] (3) We repeat this step using the most recently added word until the size of the random control set matches the size of the positive class of the original set. (4) We use the remaining words as negative control examples. The resulting set is then divided randomly into a train and test split. We use this process to create 10 control sets (each starting with a different random seed word). Results are presented in terms of mean performance scores over the 10 control sets.

Diagnostic data

Beyond a control task, the architecture of the diagnostic dataset itself can help to distinguish property learning from memorization or correlation. Ideally, examples can only be classified correctly if the classifier could identify property-specific information. Such a scenario can be approximated by a distribution which has a high diversity of examples within a class and high similarity of instances across classes. For instance, the positive examples of **fly** should consist of a large variety of concepts associated with the property (e.g. birds, insects, vehicles).

A challenging distribution allows for a targeted instance-based analysis. In particular, we can consider positive examples whose nearest neighbor in the dataset is a negative example (e.g. **fly**: *puffin* vs. *penguin*) and vice-versa. The classifier behavior on such examples can indicate whether a classifier could identify property information or whether it relied on other signals (i.e. similarity to memorized examples, correlating information). We can investigate this on test examples whose nearest training set neighbor is a member of the opposite class.

In addition, the example distribution described above increases the power of the control task. We can only expect a high difference in performance between the target and control classifiers if the target classifier has to go beyond general similarity to training set examples to solve the target task. The power of the control task thus depends on the distribution of the target data.

We use the full version of the diagnostic dataset (introduced in Part III). The dataset consists of 21 properties with positive and negative example concepts. Where possible, positive examples are taken from a diverse range of semantic categories. Negative examples have been selected in such a way that they are similar to positive examples. Table 8.5 shows the distribution of positive and negative examples and the set sizes when only considering in-vocabulary words of the embedding models we use (details in Section 8.3.3). We split each set into training and test (60%-40%).

8.3.2 Ceiling Task

We introduce a **ceiling task** to gauge the performance a classifier should be able to achieve on representations that carry information about a semantic property given a particular dataset size and distribution. We use words with a clearly marked female gender (positive class) (e.g. *nun*, *actress*) and words with clearly marked male gender (e.g. *actor*, *prince*) or without gender marking (negative class) as a ceiling dataset. Experimental results show that gender is well encoded in distributional representations (Gonen and Goldberg, 2019, among others). We use a gender dataset to mimic the size and class distributions of the property sets. If a

[12] We used the Googlenews embeddings for this step (see Section 8.3.3).

	full			voc.		
property	pos	neg	tot.	pos	neg	tot.
square	90	22	112	87	21	108
warm	133	36	169	124	32	156
black	90	53	143	78	45	123
red	92	69	161	87	64	151
fly	65	104	169	44	88	132
dangerous	77	60	137	65	51	116
wings	82	84	166	58	76	134
sweet	99	64	163	90	62	152
hot	103	43	146	100	43	143
used_in_cooking	106	65	171	100	54	154
juicy	92	64	156	84	59	143
green	94	69	163	89	66	155
made_of_wood	100	45	145	78	33	111
blue	60	110	170	59	106	165
yellow	43	88	131	43	74	117
roll	55	42	97	51	33	84
cold	70	24	94	68	24	92
round	103	20	123	96	18	114
wheels	78	27	105	69	25	94
lay_eggs	75	70	145	33	56	89
swim	101	47	148	79	38	117

Table 8.5: Size and class distribution in of the property sets in the diagnostic dataset (number of words in the full set (full) and number of words present in all model vocabularies (voc.)).

classifier achieves high selectivity on the ceiling set, we assume that information is learnable given the distribution of the property set.

To create such a dataset, we extended the **female** set created by Gladkova et al. (2016) with a manually verified selection of hyponyms of the Princeton WordNet synset of *person*. We randomly subsampled from the full set to mimic the class distribution of each property set. As with the property sets, we only include words present in all three model vocabularies. The **female** dataset did not contain enough positive examples in all model vocabularies to mimic the distributions of **used_in_cooking** and **warm**. We created ceiling sets mimicking the class distributions of all other properties.

8.3.3 Experimental Setup

In this section, we outline the components of our experimental setup: Firstly, we introduce the embedding models whose representations are used as input for the probing classifiers. Secondly, we describe the probing classifiers we use. Thirdly, we present commonly used baselines for diagnostic classification. We use these baselines to show that they cannot provide the same insights as a control task. Finally, we outline a validation procedure for the control task.

name	type	activation function	hidden layers	nodes per layer
lr	logistic regression	-	-	-
mlp1	multi-layer perceptron	relu	1	50
mlp2	multi-layer perceptron	relu	2	50, 50
mlp3	multi-layer perceptron	relu	1	100
mlp4	multi-layer perceptron	relu	2	100, 100

Table 8.6: Overview of probing classifiers.

Embedding models

We experiment with three pre-trained distributional models with the same architecture but trained on different corpora. They use the Continuous Skipgram with Negative Sampling (SGNS) algorithm suggested by Mikolov et al. (2013b). We experiment with a model trained on the full Wikipedia dump of 2017 (henceforth wiki) and a model trained on the Gigawords corpus (henceforth giga) [13] and the GoogleNews model (henceforth google).[14] The main rationales behind the model selection were to use (1) well-performing representations that are widely used (google and wiki) and (2) models whose underlying corpora are available for further exploration in future work (wiki and giga, see corpus analysis in Chapter 7). We expect differences in terms of property expression between Wikipedia texts and news texts as they can be expected to emphasize different aspects of conceptual information.

Probing Classifiers

We experiment with a linear regression model (lr) and four multi-layer perceptrons (mlp). The mlp models use a relu activation function and have 1 or 2 hidden layers consisting of either 50 or 100 nodes (for details, see Table 8.6). We did not optimize the settings. We report the mean performance over ten runs for each mlp probe.

Baselines

We compare against the following, simple baselines:

- random label assignment
- randomly initialized vectors (mean performance over 10 sets each)
- majority class assignment

We use the baselines to establish that the control task poses a considerably higher bar.

[13] Both trained following recommended settings by Levy et al. (2015). The models can be downloaded from: `https://bitbucket.org/PiaSommerauer/distributionalmodels`.

[14] Downloaded from `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit`.

Control Tool Validation

We use several checks to establish that our control tools (i.e. the challenging example distribution in the property sets and the quality of the control task) used to interpret the diagnostic classification results can provide meaningful results.

Firstly, we check the differences between the vector similarity distributions of the property datasets. Ideally, the property sets should feature high example diversity within class (i.e. low average similarity within class containing examples such as *seagull*, *bee airplane* for the property **fly**) and high similarity across classes (e.g. **fly**: *puffin* vs. *penguin*). We calculate mean pairwise cosine similarity among the words within the positive class and across the positive and the negative class of a property. Overall, we can expect that a datasets with high across-class and low within-class similarity will contain more challenging across-class pairs with high similarity.

Secondly, we check whether the control task is a sufficiently high bar. We establish this by comparing the performance of the probing classifiers on the control task to the baselines mentioned above. The probing classifiers should clearly perform higher on the control task than on the baseline tasks.

Thirdly, we check whether the control task sets are sufficiently different from the property task sets. The similarity chain we use to form the control sets has the risk of assigning a large proportion of positive or negative property examples to the same class in the control task. Consider this (exaggerated) scenario: All positive examples of the property **red** are red fruits and all negative examples are green vegetables. If we pick a random example to populate our control set (e.g. *strawberry*), it is likely that its nearest neighbor comes from the same class (e.g. *raspberry*). The same risk applies if the first random word comes from the negative class (e.g. *broccoli*). Consequently, the control task would be almost equivalent to the property task and thus become less informative. We check how many examples in the positive class in the control set come from the same class (positive or negative) in the original set.

8.3.4 Results

We first validate the suitability of our diagnostic dataset and the control task. We then present our diagnostic classification results.

Control tool validation

Similarity distribution For each property, we show the difference between the mean positive-class-similarity and the mean across-class-similarity (summarized in Table 8.7). A value below zero indicates that the across class similarity is, on average, higher than the within class similarity (i.e. the property set is challenging). A value above zero indicates that the examples in the positive class are, on average, more similar to one another than to examples of the negative class (i.e. the property set is less challenging). All three embedding models yield similar results. For the first five properties, across class similarity is higher than within class similarity or at least equally high (**red**, **round**, **roll**, **blue**). This can be expected, as the properties constitute perceptual attributes that tend to cut across different semantic categories (e.g. colors can apply to a large variety of artifacts, but also natural things). For all

other properties, the similarity within the positive class is higher. While the differences are still rather low for most properties, we can observe that they are highest for properties that tend to correlate with taxonomic categories (**wheels**, **wings**, **used_in_cooking**).

properties	wiki	giga	google
red	-0.01	-0.01	-0.01
round	-0.01	-0.02	-0.02
black	0.00	-0.00	0.00
blue	0.00	-0.00	-0.01
made_of_wood	0.00	0.01	0.02
roll	0.01	-0.01	-0.00
yellow	0.01	-0.00	0.01
juicy	0.02	0.03	0.07
green	0.02	0.03	0.06
cold	0.03	0.03	0.04
sweet	0.03	0.03	0.05
swim	0.03	0.05	0.05
warm	0.03	0.03	0.03
hot	0.04	0.05	0.05
fly	0.05	0.02	0.06
dangerous	0.05	0.04	0.07
wheels	0.06	0.05	0.09
square	0.06	0.07	0.10
lay_eggs	0.06	0.06	0.11
wings	0.08	0.04	0.15
used_in_cooking	0.13	0.19	0.19

Table 8.7: Difference between the average pairwise similarities within the pos. class and across classes (pos - pos_neg). Negative values indicate that the similarity across classes is higher than the similarity within the positive class.

Control task performance All probes perform considerably higher on the control task than on the baselines (Table 8.8), indicating that the control task is an easy task and therefore provides a sufficiently challenging control tool.

Overlap between property task and control task On average, the control set does indeed contain a considerable proportion of examples from a single original class (70% in the property sets and 82% in the ceiling sets). However, a dataset containing 20% to 30% of noise can be expected to considerably hamper performance, in particular given the relatively small size of the individual sets. The degree of overlap differs between properties, as can be observed in Table 8.9.

Probing analysis

We first analyze the results using selectivity and then explore particularly challenging examples.

control	lr	mlp1	mlp2	mlp3	mlp4
wiki	0.76	0.75	0.76	0.75	0.76
google	0.82	0.81	0.82	0.81	0.81
giga	0.78	0.77	0.78	0.77	0.78
random labels					
wiki	0.12	0.13	0.12	0.13	0.13
google	0.12	0.13	0.13	0.13	0.13
giga	0.12	0.13	0.12	0.13	0.13
random vectors					
wiki	0.57	0.57	0.57	0.57	0.57
giga	0.57	0.57	0.57	0.57	0.57
google	0.57	0.57	0.57	0.57	0.57
			majority		
all			0.53		

Table 8.8: Mean performance (weighted f1 score) on control set and baselines (random labels and vectors, and majority).

property	prop-set	ceiling
red	0.54	0.86
square	0.88	0.99
fly	0.75	0.65
black	0.56	0.94
used_in_cooking	0.81	0.94
wings	0.69	0.67
made_of_wood	0.74	0.98
hot	0.77	0.94
sweet	0.63	0.87
green	0.62	0.82
yellow	0.57	0.62
juicy	0.65	0.79
blue	0.76	0.55
roll	0.59	0.83
lay_eggs	0.58	0.56
round	0.83	0.98
cold	0.75	0.96
swim	0.74	0.91
wheels	0.78	0.94
dangerous	0.69	0.78
mean	0.70	0.82

Table 8.9: Example overlap in control datasets.

Selectivity Figure 8.1 shows the probing results (F1) and selectivity scores (D) on the wiki model for the property (p) and ceiling sets (c). All available ceiling sets show a selectivity

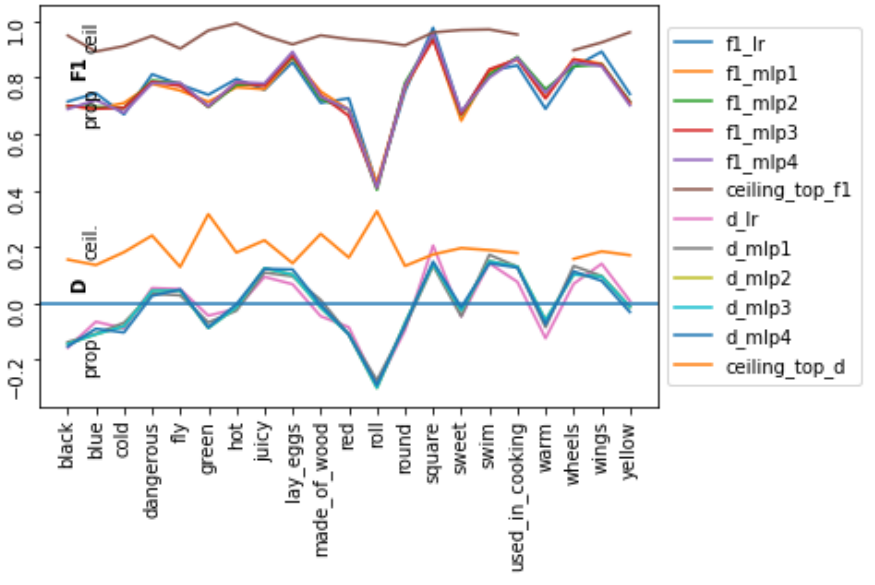


Figure 8.1: Weighted f1 score '**F1**' and selectivity '**D**' of the property (p) and ceiling task (c) in the wiki model.

d	giga	wiki	googlenews
> 0	**fly juicy lay_eggs square** sweet swim **used_in_cooking wheels wings**	dangerous **fly juicy lay_eggs** made_of_wood **square** swim **used_in_cooking wheels wings** yellow	**fly juicy lay_eggs square** sweet **used_in_cooking wheels wings**
< 0	**black blue cold** dangerous **green hot** made_of_wood **red roll round warm** yellow	**black blue cold green hot red roll round** sweet **warm**	**black blue cold** dangerous **green hot** made_of_wood **red roll round** swim **warm** yellow

Table 8.10: Properties with positive and negative selectivity values in the three models (giga, wiki, google).

clearly above 0 in the wiki embeddings. For clarity, we only show the top classifier.[15] Figure 8.1 includes all five classifiers (logistic regression and 4 multi-layer perceptrons) for properties, showing that they achieve similar performance and selectivity. We observe that high performance does not necessarily mean high selectivity, but generally seems to correlate with it. For the property **roll**, the difference between property set and ceiling set is particularly stark (for selectivity and f1 score). In this case, the difference between ceiling and property set is likely to indicate that information is learnable given the dataset size and class distribution (indicated by high ceiling performance and selectivity), but the property-set does not seem

[15]The google embeddings do have one ceiling set score that remains below 0: the one mimicking **green**.

to capture learnable property information. The analysis of the property datasets presented in Chapter 7 showed that the property set for **roll** is indeed likely to be particularly difficult for classifiers and to contain noise. Detailed results of all models and different classifiers are provided in the appendix.

property	diff tot.	✓ abs.	acc.	same tot.	✓ abs.	acc.
square*	3	3	1.00	39	38	0.97
used_in_cooking*	16	14	0.88	46	40	0.87
lay_eggs*	7	6	0.86	22	20	0.91
swim*	9	6	0.67	36	32	0.89
juicy*	20	13	0.65	37	29	0.78
wings*	5	3	0.60	41	38	0.93
yellow*	12	7	0.58	32	26	0.81
green	21	12	0.57	36	30	0.83
red	29	16	0.55	31	28	0.90
wheels*	6	3	0.50	32	30	0.94
dangerous*	14	7	0.50	28	27	0.96
round	9	4	0.44	28	27	0.96
made_of_wood*	7	3	0.43	30	25	0.83
blue	19	8	0.42	41	37	0.90
fly	12	5	0.42	36	33	0.92
hot	16	6	0.38	38	37	0.97
warm	15	5	0.33	45	45	1.00
black	13	4	0.31	31	27	0.87
sweet	14	4	0.29	47	37	0.79
roll	13	2	0.15	15	11	0.73
cold	7	1	0.14	25	24	0.96

Table 8.11: Number (n) and proportion (acc.) of correctly classified examples with different class nearest neighbors (diff) and same class (same) nearest neighbors in the wiki model. Properties marked with * have a positive selectivity score.

Properties The selectivity scores are generally lower for the property tasks than for their respective ceiling tasks. For several properties, the selectivity scores are only marginally above 0. Table 8.10 shows the properties for which the probing classifiers obtained a selectivity score above 0. When considering the successfully learned properties, it should be considered that the properties **lay_eggs** and **used_in_cooking** (successfully classified for all models) run a particular risk of containing accidental correlations (see analysis in presented in Chapter 7).

The different distributional models have the same outcome for 16 out of 21 properties. The difference for the other 5 properties may be explained by the different genres they are trained on. In contrast to the two new-based models, wiki seems to encode the properties **dangerous**, **made_of_wood**, and **yellow**. It could be argued that these three properties are more likely to be mentioned in descriptive, encyclopedic texts than in news texts. The property **yellow** is the only color property with a selectivity score above 0, indicating that colors are not well represented in corpus data.

property	diff tot.	✓ abs.	acc.	same tot.	✓ abs.	acc.
used_in_cooking*	10	9	0.90	52	49	0.94
lay_eggs*	7	6	0.86	23	21	0.91
sweet*	16	12	0.75	45	36	0.80
round	13	7	0.54	25	25	1.00
blue	21	11	0.52	39	34	0.87
hot	16	7	0.44	39	35	0.90
yellow	17	7	0.41	28	24	0.86
swim*	10	4	0.40	35	31	0.89
warm	13	5	0.38	50	45	0.90
fly*	11	4	0.36	37	34	0.92
red	28	10	0.36	33	29	0.88
wings*	6	2	0.33	40	36	0.90
roll	18	6	0.33	10	7	0.70
green	16	5	0.31	42	33	0.79
made_of_wood	11	3	0.27	26	24	0.92
black	15	4	0.27	29	27	0.93
wheels*	5	1	0.20	33	29	0.88
square*	5	1	0.20	37	37	1.00
juicy*	13	2	0.15	45	37	0.82
cold	9	1	0.11	24	23	0.96
dangerous	10	1	0.10	33	31	0.94

Table 8.12: Number (n) and proportion (acc.) of correctly classified examples with different class nearest neighbors (diff) and same class (same) nearest neighbors in the giga model. Properties marked with * have a positive selectivity score.

The properties with selectivity scores below 0 for all models constitute visual-perceptual properties, inline with hypotheses and observations from previous work that this information is probably not encoded. Properties with positive scores encompass a variety of property types. As also found in previous work, several properties with positive selectivity for all models correlate with taxonomic categories (**wheels**, **used_in_cooking**, **wings**).

Challenging examples A classifier that learned to identify a property should be able to distinguish nearest neighbors that belong to an opposite class. For instance, the word *rabbit* is the training example closest to test example *duck* according to the wiki embeddings. If a classifier can identify the property **fly**, it should be able to recognize *duck* as a positive example, despite its high similarity to the negative example *rabbit*. If a classifier generally fails at identifying such different-class examples, it probably did not learn to identify property-specific information. Instead, it may have relied on correlations (e.g. information about a semantic category).

We compare classifier accuracy on different-class pairs ('diff') to nearest neighbor pairs that belong to the same class ('same') in wiki (Table 8.11) and giga (Table 8.12). Unsurprisingly, scores are consistently higher for same-class examples (between 0.95 and 1 in wiki and 0.70 for giga). For 10 out of 21 properties, the accuracy of different class examples is below

property	diff tot.	✓ abs.	acc.	same tot.	✓ abs.	acc.
roll*	7	7	1.00	21	19	0.90
lay_eggs*	5	5	1.00	25	23	0.92
hot*	8	8	1.00	47	47	1.00
swim*	11	10	0.91	34	34	1.00
yellow*	10	9	0.90	35	34	0.97
sweet*	19	17	0.89	42	41	0.98
green*	16	14	0.88	42	42	1.00
wings*	15	13	0.87	31	29	0.94
juicy*	19	16	0.84	39	39	1.00
made_of_wood*	6	5	0.83	31	30	0.97
dangerous*	11	9	0.82	32	31	0.97
black*	11	9	0.82	33	33	1.00
used_in_cooking*	10	8	0.80	52	50	0.96
round*	7	5	0.71	31	30	0.97
red*	17	12	0.71	44	42	0.95
wheels*	9	6	0.67	29	28	0.97
square*	6	4	0.67	36	36	1.00
blue*	16	10	0.62	44	43	0.98
fly*	7	4	0.57	41	39	0.95
cold*	6	3	0.50	27	27	1.00

Table 8.13: Number (n) and proportion (acc.) of correctly classified examples with different class nearest neighbors (diff) and same class (same) nearest neighbors in the wiki model for the ceiling sets. Ceiling sets marked with * have a positive selectivity score.

0.50 in wiki and for 16 out of 21 in giga. The highest score in giga (1.00) is achieved for the property **square**. The low number of different-class nearest neighbor examples in combination with the relatively high within-class similarities may be an indication that the dataset contained accidental correlations that help to distinguish positive from negative examples without detecting property-specific information. In giga, **square** scores considerably lower (0.20). The second highest score in wiki (0.88) is achieved for **used_in_cooking** (based on 16 examples), but also high similarity within the positive class. In giga, the highest score is achieved for **used_in_cooking** (0.90 based on 10 examples).

In wiki, it can be observed that properties with a high selectivity score tend to show relatively high performance on different-class examples. Seven of the ten properties with positive selectivity score achieve an accuracy above 0.50, three score 0.50 or below. Most properties with a selectivity score below 0 also score very low on different class examples. In giga, this tendency can also be observed, but not quite as strongly. The three best performing properties with a score of at least 0.75 also have selectivity scores above 0. The remaining properties with selectivity scores of above 0 score below 0.50. In contrast, the corresponding ceiling mimicking the property set distributions sets yield comparatively high scores (shown for wiki in Table 8.13): For all but one ceiling set, the classifiers reach scores above 0.50 for different class examples. 13 sets score 0.80 or above.

We explore the different-class nearest neighbor pairs for the properties shown in Table 8.14

(wiki) and Table 8.15 (giga). Next to high semantic similarity or relatedness between examples (e.g. *deer-pheasant*, *rabbit-duck*), we observe the following factors that are likely to be challenging for a probing classifier:

- polysemous examples (*bass*, *club*, *hack*)
- instances that are only weakly associated with the property (**fly**: *toy*, *machine*; **swim**: *mammal*)
- annotation inconsistencies: some mammals that can, but usually do not **swim** are treated as positive examples, some as negative examples (*lion*-neg *wolf*-pos, *deer*-pos, *llama*-neg)
- vague examples: *window* can refer to a glass cover or a wooden frame around it

Though a successful model should be able to identify polysemous examples (unless the intended meaning is very rare), these vague examples could be instances that do not encode the property even though other instances do. Correctly classified pairs also include polysemy (e.g. **lay_eggs**: *duck-cock*) and classifications of challenging examples related to fine-grained subcategories **wheels**: *passenger-luggage*, **lay_eggs**: *whale-albatross*). Based on this small number of examples, we suspect that the models can pick up relatively fine-grained taxonomic classes, but do not necessarily learn the isolated properties. To confirm this, a larger and more systematic analysis would be required.

To summarize, the results of the ceiling task provide strong indications that probes can indeed identify information given small and skewed data distributions if the information in question provides a strong signal. For gender marking, selectivity scores are high and challenging examples tend to be classified correctly. For the property datasets, however, we could not identify equally strong signals. Previous work suggests that embeddings (partially) encode taxonomic property knowledge, but not visual properties. Our results are in line with this as they reveal weak positive selectivity scores for taxonomic properties and below zero selectivity scores for visual properties. Low performance on challenging examples leads us to suspect that classifiers pick up evidence of fine-grained taxonomic structures rather than isolated semantic properties.

8.3.5 Conclusion

In this study, we applied a probing analysis of lexical embedding representations with respect to semantic properties. We addressed the known problem of probing that imperfect above-chance performance is difficult to interpret by applying two control analyses (selectivity and challenging examples) and a ceiling task (gender encoding).

Combining vanilla probing with these two control tools and the ceiling tasks led to the following insights on probing in general and identifying semantic properties in particular. In general, our results show that probing, in combination with control tasks, can indeed detect information in embedding representations if there is a strong signal. Gender information is encoded systematically in co-occurrences. This is reflected in high selectivity scores and high accuracy on challenging examples, even on relatively small and imbalanced datasets.

property	correct	incorrect
fly	deer-quail* penguin-puffin* flycatcher*-watchdog loon*-van gull*-barnacle	tractor-machine* rabbit-toy* interceptor-twinjet* interceptor-bomber* interceptor-jet* rabbit-duck* deer-pheasant*
dangerous	hornbill-hippopotamus* rabbit-coyote* malaria*-cure killer*-jack hammer*-hoe alligator*-turtle snake*-toad	starter-club* crowbar-punk* tuck-raper* crowbar-nightstick* bootlegger*-businessman sword*-elixir shaft*-pusher
lay_eggs	whale-albatross* cow-chicken* whale-leatherback* rattlesnake*-howler duck*-cock crocodile*-giraffe	rattlesnake*-mole
made_of_wood	strap-pin* knife-toothpick* strap-footrest*	pulley-shaft* window*-date pencil*-pen bass*-pedal
swim	bat-runner* chicken-duck* lion-wolf* roach-guppy* albatross*-cockatoo deer*-llama	squirrel-armadillo* mammal*-bonobo mammal*-colobus
wheels	driver-coach* winch-rig* passenger-luggage*	backhoe*-hack locomotive*-diesel cab*-windshield

Table 8.14: Examples of correctly and incorrectly classified examples with different-class nearest neighbors in the training set for the wiki embeddings. The first example always corresponds to the training example, the second to the test example. Words marked with * are positive examples of the property.

For semantic properties, we learned that imperfect performance on the target set cannot be (solely) attributed to skewed datasets, since the ceiling sets mimic these distributions. Secondly, the selectivity analysis shows that the probing signal picked up for several taxonomic categories goes beyond general similarity, while it confirms the result that visual-perceptual information is not encoded. Thirdly, our analysis on challenging examples indicates that, nevertheless, these results do not seem to be the result of identifying isolated properties, but rather of learning fine-grained categories. Overall, these results are inline with previous work stating that taxonomic properties are (partially) present and visual properties are absent. Future work with more systematic error analysis (notably concerning vague examples) could shed a light on whether learning fine-grained categories is indeed the explanation for these results.

Overall, we conclude that our methodological controls (selectivity, challenging examples and the ceiling task) together considerably improve the interpretation of probing results. In particular, challenging examples allowed us to go beyond selectivity showing that weak positive signals do not seem to indicate isolated property knowledge after all. The ceiling task enabled us to rule out that this result is solely due to the size and class imbalance. Nevertheless, we do not have conclusive answers for weak signals. Future work will have to show whether further data analysis can address this, or whether probing reaches its limits

property	correct	incorrect
fly	penguin-puffin* loon*-chow budgie*-rudd pelican*-barnacle	tractor-machine* rabbit-toy* interceptor-arrow* zebra-marabou* penguin-nightingale* deer-pheasant* deer-quail*
dangerous	iodine-neurotoxin*	bar-club* rabbit-coyote* blunderbuss*-bill malaria*-cure thief*-businessman thief*-pusher knife*-hoe alligator*-turtle snake*-toad
lay_eggs	whale-leatherback* crane*-hack platypus*-howler crane*-hackney halibut*-mole crocodile*-giraffe	hobby-lark*
made_of_wood	wheelbarrow-chock* wheel-clock* knife-toothpick*	saxophone-guitar* chimney-shaft* window*-date stool*-anchor pencil*-pen dowel*-screw roof*-windshield broom*-trowel
swim	chicken-cob* roach-loach* painter*-sweeper retriever*-cockatoo	chicken-goat* chicken-duck* mammal*-bonobo deer*-llama mammal*-colobus frog*-owl
wheels	passenger-luggage*	tugboat-rig* passenger-airplane* machine*-hack car*-windshield

Table 8.15: Examples of correctly and incorrectly classified examples with different-class nearest neighbors in the training set for the giga embeddings. The first example always corresponds to the training example, the second to the test example. Words marked with * are positive examples of the property.

here.

8.4 Summary

In this chapter, I have presented two diagnostic studies that aim to test whether context-free embedding representations contain information about different semantic properties. Both studies approach the challenges of diagnostic classification through control tools. The purpose of these tools is to distinguish high probing performance caused by property-identification from high probing performance caused by accidental correlations (e.g. in an extreme scenario, all positive examples of **fly** are in the category BIRD, all negative examples in the category FURNITURE) or memorization (meaning that the classifiers simply assign labels based on general similarity to training examples).

The results of the first study provided the following, preliminary indications: (1) Visual-perceptual properties achieved low performance in the probing task, which indicates that they are most likely not encoded well in embedding vectors. (2) The results provided some indications that properties connected to activities and functions may be encoded in embedding representations (in particular for the properties **used_in_cooking**, **wheels**, and **dangerous**.

However, the study is limited by the fact that the pilot version of the diagnostic datasets are likely to contain correlations (in particular with taxonomic categories). The comparison to nearest neighbor classification is not necessarily fine-grained enough to eliminate the possibility that the classifiers simply relied on such factors.

The second study approaches the methodological problems of diagnostic classification by exploiting the challenging example distribution of the full diagnostic dataset and a semantic control task. The semantic control task poses a challenging baseline: Classifiers can achieve high performance on it by relying on general semantic similarity without identifying the target information in the embeddings. The challenging distribution of examples allows for a targeted error analysis which can provide insights into whether probing classifiers could distinguish highly similar examples that differ with respect to the target property. In addition, the study employed a ceiling task which established determine whether information is learnable given small datasets with skewed class distributions.

The results of the control task clearly indicate that overall, property-information performs much lower than gender-marking used in the ceiling datasets. Properties that did consistently perform highly tend to correlate with relatively coherent semantic categories (e.g. **used_in_-cooking**, **lay_eggs**, **wheels**, **juicy**, **fly**). The analysis of challenging examples showed that the classifiers are not well-equipped to distinguish highly similar examples with respect to the target properties (e.g. *rabbit* and *duck* could not be distinguished with respect to **fly**). The relatively low selectivity scores in combination with the low performance on challenging across-class examples may indicate that the classifiers learn to identify fine-grained semantic categories rather than property-specific information.

Despite our control tools, the following limitations remain: Firstly, despite the controlled dataset distributions, there is still a risk of accidental correlations between the examples of a class. Secondly, diagnostic classification assumes that the target information is encoded in the majority of examples in the train and test set. If the information is only encoded in a proportion of examples, the classifiers will inevitably fail or rely on other 'clues' to perform the task. To shed more light on whether property information is likely to be encoded in the property-examples, I turn to an analysis of the corpus data underlying the embedding models in Chapter 9.

9. Evidence Analysis in two Corpora

9.1 Introduction

'Diagnosing' semantic properties in diagnostic classification experiments (Chapter 8) has a central problem: Even if classifiers can successfully perform the diagnostic classification task (e.g. distinguish positive and negative examples of the semantic property **red** on the basis of embedding representations of words such as *strawberry* and *dog*), it remains difficult to determine whether the classifiers have indeed identified evidence of the target property or whether they have relied on other features that happened to correlate with the target information. The diagnostic dataset used for the experiments was specifically designed to reduce the chance of such correlations (Chapter 4). Nevertheless, the analysis of classification errors presented in the previous chapter provided reasons to question whether the classifiers have indeed identified property-specific information in the embedding vectors. Even high performing classifiers struggled with distinguishing highly similar pairs of positive and negative examples (e.g. **fly**: *duck* v.s. *rabbit*). A possible explanation of this classification behavior could be that the classifiers identify fine-grained semantic categories that correlate with many but not all examples in the datasets. This chapter presents an analysis of the corpus data underlying the embedding models to verify these findings. The corpora under consideration are the data underlying two of the three embedding models used for diagnostic experiments:

- Wikipedia full dump 2018 (henceforth wiki)
- Gigawords 5th edition (henceforth giga)

To extract candidates of property expression from the corpus data, I rely on the contrastive nature of the dataset. The positive and negative examples of each property can be used to compare the contexts of positive examples against the contexts of negative examples (e.g. **fly**: contexts of *seagull* v.s. contexts of *penguin*). If semantic properties are mentioned in the contexts of concepts, this type of contrastive analysis should highlight them, as property expressions should be represented more strongly in the context of positive examples than of negative examples. The details of the method used for evidence extraction and analysis are provided in Section 9.2.[1]

Section 9.3 presents an analysis of the extracted candidates for property-evidence. The results indicate that most properties are, at least to some degree, expressed by property-specific linguistic expressions (e.g. the property **red** is expressed by the adjective *red*). Positive examples of properties also co-occur with other concepts that share the target property (e.g.

[1]The code used for the context extraction and analysis can be found in this repository: `https://github.com/PiaSommerauer/CorpusDiagnostics`.

bird names in the dataset for the property **fly** tend to co-occur with other bird names). I refer to this type of indirect property-evidence as 'property-instances'. Positive examples also tend to co-occur with words related to the property via thematic associations (e.g. flying vehicles in the dataset for **fly** tend to co-occur with thematically related concepts such as *pilot*). I refer to this latter type of property-evidence as 'property-related words'. To gain insights into what type of information diagnostic classifiers are likely to have picked up, I analyze the relation between property evidence in corpus data and the results of the diagnostic experiments presented in the previous chapter. The results indicate that property-instances and related words are more likely to be represented strongly in the embeddings and picked up by diagnostic classifiers than property-specific evidence. This finding strengthens the hypothesis that distributional co-occurrence patterns tend to encode fine-grained semantic categories rather than semantic properties.

A second purpose of the corpus analysis presented in this chapter is to gain deeper insights into the expression of semantic property evidence in corpus data. Previous research has argued that specific types of properties are more likely to be expressed than others (e.g. taxonomic v.s. perceptual properties (Rubinstein et al., 2015)). The diagnostic dataset presented in this thesis is based on a model of the dynamics of property expression in textual data (introduced in Chapter 3). The model's fundamental assumption is that the expression of semantic properties in corpus data depends on the relation between a specific concept and the property in question. The relations between properties and concepts are based on factors that are likely to influence whether property information is made explicit. For example, highly implied information (expressed by the property-concept relation `implied_category`) is not expected to be made explicit (e.g. **mammal** - *cat*). In contrast, information that is relevant for how we interact with the world (e.g. **cut**-*scissors*) is expected to be mentioned. In addition to property types and property-concept relations, it can be expected that the genre of a corpus also impacts what type of conceptual information is expressed. For instance, encyclopedic texts can be expected to be more explicit about highly implied knowledge, while news texts can be expected to emphasize events (which likely mention afforded actions).

Section 9.4 presents an analysis of property evidence with respect to property-concept relations (the central component of the theoretical framework presented in Chapter 3), genre, and property types. As explained in Chapter 7, the analysis of property-concept relations is complicated by interactions between different relations. To address this issue, I use two complementary strategies of assigning property-concept pairs to specific relations. The results provide initial indications that are, at last partially, in line with the hypotheses derived from the theoretical framework presented in Chapter 3. It seems that property evidence is likely to be mentioned if a property affords a specific activity (e.g. **having sharp edges** affords cutting with *scissors*). Initial tendencies also indicate that variability between properties (e.g. an *apple* can be **red**, **green**, or **yellow**) may lead to more explicit property expressions. The latter tendency has also been observed in contemporary research about the reporting bias (Paik et al., 2021). The analysis on the level of individual properties showed weak tendencies in line with previous research: Overall perceptual properties seem to be expressed less strongly than other properties. However, this trend is not consistent across all properties. This inconsistency is expected according to the hypotheses presented in Chapter 3.

9.2 Data and Method

In this section, I provide an overview of the methods used for context extraction and analysis. Section 9.2.1 outlines the extraction of concept-contexts from corpus data. In a next step, I extract property-evidence candidates from the contexts by means of a contrastive analysis of positive and negative property-examples (Section 9.2.2). I annotate the extracted evidence-candidates with respect to whether they can be seen as evidence of the respective semantic property under investigation (Section 9.2.3). Based on the annotated evidence, I employ different measures to analyze the degree to which property evidence is represented among the examples of a property (Section 9.2.4).

9.2.1 Corpora and Context Extraction

To get insights into the information distributional models can exploit, I analyze the contexts which are used to train word embedding models. In the standard set-up of training a context-free distributional model, word-context pairs are extracted from a pre-processed corpus. I use the same extraction process as the one used by the Word2vec implementation of the skip-gram with negative sampling (SGNS) method to ensure that the data under investigation reflect the same information as the models used in the diagnostic classification experiments.

The word-context extraction is dependent on a number of hyper-parameters of the model. The models under consideration were trained using the recommended settings based on an analysis of hyper-parameters by Levy et al. (2015). The most important setting for the word-context extraction from the raw data is the window size. As recommended, I used a window size of 2 (meaning that the model 'sees' two words left and right of the a target word). In addition to window size, the following factors also affect the creation of word context pairs: The SGNS algorithm uses sub-sampling to remove highly frequent words from the training data. The method also removes rare words. Both sub-sampling and removing rare words are done before the word-context pairs are created.

9.2.2 Extraction of property-evidence candidates

It is hardly feasibly to analyze all linguistic contexts of all positive examples of a property by hand. Instead, I employ the following criteria to capture candidates that are likely to function as property-evidence:

1. Contexts should be particularly salient in the positive examples of a property when compared to the contexts of its negative examples.

2. Contexts should be particularly good at distinguishing positive examples of a property from negative examples.

In addition to limiting the scope of the analysis, prioritizing salient and distinctive contexts reduces the chances of noise. Salience and distinctiveness are defined below.

Salient Contexts

To identify contexts that are particularly salient in the contexts of positive examples of a property, I use a frequency measure commonly applied in information retrieval. The original goal of the measure is to identify terms that are particularly salient in a specific document. For example, such salient terms can be used to match documents with search terms. To find salient words for a given document, the frequency of a term t within a document d is compared to the number of documents N it appears in (out of all documents D). Words that are frequently used within specific document, but only appear in comparatively few documents can be seen as characteristic of the document and thus receive a high value. This notion has been formalized as term frequency-inverse document frequency (tf-idf) (Spärk Jones, 1972). Term frequency is defined as the normalized frequency of a word in a document:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{9.1}$$

Inverse document frequency is defined as:

$$idf(t,D) = log \frac{N}{|\{d \in D : t \in d\}|} \tag{9.2}$$

Term frequency - inverse document frequency then becomes:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{9.3}$$

I use this intuition to find contexts of positive examples of a property that are particularly salient when compared against contexts of negative examples of a property. If the property is expressed, the salient contexts should include expressions of the property. For example, when comparing the contexts of the word *strawberry* against the contexts of negative examples of the property **red** (e.g. *pineapple*, *orange*, *mango*, *banana*), the word *red* should be among the most salient contexts of *strawberry* if it is mentioned explicitly and consistently. I thus adapt the tf-idf definition to measure the salience of a context word for a particular target word as follows:

Context frequency. All context words of a target word (i.e. a positive example of a property) are treated as a single document. Together, the contexts represent the target word w. To calculate the frequency of a particular context word c for a particular word w, I simply apply term frequency as follows to calculate 'context frequency' (cf):

$$cf(c,w) = \frac{f_{c,w}}{\sum_{c' \in w} f_{c',w}} \tag{9.4}$$

Inverse word frequency. To contrast the relevance of the context word for the target word against the negative examples, I treat the context words of each negative example as a document w_{neg}. Together, they form a collection of negative example words W_{neg}. I count the number of times N the context word appears in the contexts of a negative example and calculate 'inverse word frequency' as follows:

$$iwf(c,W_{neg}) = log \frac{N}{|\{w_{neg} \in W_{neg} : c \in w_{neg}\}|} \tag{9.5}$$

Based on context-frequency and inverse word frequency, I calculate the $cf\text{-}iwf$' score:

$$cfiwf(c, w, W_{neg}) = cf(c, w) \cdot iwf(c, W_{neg}) \tag{9.6}$$

I use this definition to calculate the cf-iwf score of each context word of each positive example concept and keep all contexts with the top 10000 cf-iwf scores.[2] I also calculate the cf-iwf scores for context words of negative examples. For contexts of negative examples, the calculation is simply flipped: Each negative example concept is treated as a document and the contexts of all positive examples are used as the corpus for comparison.

The cf-iwf values of contexts of positive and negative examples can be used to establish whether a context word is particularly salient for a given positive example of a property. I use the following inclusion criteria to collect salient contexts: (1) I include all words that do not appear as context words of negative examples. (2) Out of the words which do appear as context words of negative examples, I include words that have a higher cf-iwf value than their mean cf-iwf value calculated over all negative examples.

Within-semantic category comparison. A central characteristic of the diagnostic dataset is that the examples of properties are sampled from different semantic categories. For example, the positive and negative examples of the property **fly** contain, among others, words in the categories BIRD, VEHICLE, and FOOD. When comparing a positive example from the category BIRD against all negative examples, the cf-iwf measure runs risk of identifying contexts that are specific to the semantic category rather than the property **fly**. Therefore, I limit the comparison against positive/negative contexts to words within the same semantic category as the target word. For example, I calculate cf-iwf scores for the contexts of the word *seagull* by means of comparing it to all negative examples in the category BIRD instead of the full range of negative examples.

To exploit the semantic categories represented in the dataset to their full extent, I make use of the original WordNet synsets used to retrieve candidates for the diagnostic dataset (see Chapter 4 for details). Specifically, I attempt to sort all positive and negative examples retrieved via a source that did not supply category information (e.g. the semantic space or property norms) according to the synsets via their WordNet synsets and hyponymy relations. Only words for which I cannot find a semantic category remain in their original, general category.

Despite controlling for the semantic categories, high cf-iwf scores still indicate contexts that are highly specific to a particular example concept. For instance, when considering the property **fly** in the giga corpus, the top context of the word *nightingale* is *Florence* and the fourth most specific context of the word *seagull* is *Jonathan*[3]. Words that can act as potential property evidence tend to be ranked highly, but not necessarily on top (e.g. *flies* appears on rank 13 for *seagull*). Context salience alone is not sufficient to detect good property-evidence candidates. I use an additional measure to select relevant evidence words from all salient context words.

[2] I use the tf-idf implementation in Scikit Learn (Pedregosa et al., 2011).
[3] Novel: Jonathan Livingston Seagull by Richard Bach

Distinctive Contexts

Good property-evidence candidates should serve as a distinguishing feature between positive and negative examples. For instance, the word *red* expressing the property **red** should only co-occur with positive examples. I use the salient evidence candidates to find the most distinctive contexts. I measure distinctiveness as follows: I use a particular context word to distinguish positive from negative examples of a property. For instance, for the property **red**, I consider the context word *red* and calculate how well it distinguishes positive (e.g. *strawberry*) from negative examples (e.g. *grass*). I calculate the performance of the context word in terms of its precision and recall (summarized by the f1-score). I select the context words with the top three f1 scores per semantic category for further analysis. This results in substantial sets of property-evidence candidates (shown in Table 9.1), as multiple words can have the same score and many property datasets have concepts in a variety of different semantic categories.

	giga	wiki
square	1504	441
warm	2108	1181
black	1149	819
red	1767	1275
fly	952	55
dangerous	1114	627
wings	560	332
sweet	35	26
hot	84	92
used_in_cooking	572	705
juicy	810	122
green	578	550
made_of_wood	786	497
blue	2234	1818
yellow	52	111
roll	3886	3485
cold	1081	590
round	520	1034
wheels	244	105
lay_eggs	74	26
swim	1381	1700
female (control)	45	109

Table 9.1: Overview of evidence candidates per property in the giga and wiki corpus.

To validate the context selection strategy based on salience and distinctiveness, I present an overview of the extracted contexts from the giga corpus in Table 9.2. The table shows the total number of extracted contexts, the f1 score of the top-ranked contexts and the top contexts themselves.[4] Several contexts either express a property directly or are closely related

[4]The top f1 score is calculate on the basis of the mean f1-scores over all semantic categories with at least ten positive and negative examples.

to it (e.g. **red**, **made_of_wood**, **dangerous**, and **female**). For other properties, the top distinctive contexts are entirely unrelated (as can be observed for the properties **swim** and **warm**). This outcome is expected, as some properties may have fewer evidence expressions in the corpus. As in the experiments presented in Chapter 8, the property **female** is used as a control. The fact that sensible context words could be extracted for the control property is a strong indication that the extraction method based on salience and distinctiveness yields highly relevant candidates for property-evidence. This is also confirmed by other examples in the table. The total sets of evidence candidates (i.e. all extracted evidence candidates) are used for further analysis.

property	f1-mean	contexts
used_in_cooking	0.94	add recipe fish
fly	0.90	payload study experimental hovering flew overhead
green	0.86	place shade belongs green citrus
square	0.85	built
blue	0.85	magic bright scale various european
lay_eggs	0.83	eggs
hot	0.83	flame oven remove hot
sweet	0.83	banana potato sweet
yellow	0.83	sipping operating apple yellow good bubbly apples
black	0.83	grandmother suddenly named tracking introduction fire
juicy	0.82	banana ripe pineapple for
wheels	0.82	truck driver drove wheel
dangerous	0.82	killed
cold	0.82	fresh variety contains
roll	0.79	35 half
red	0.79	red intense wine currant burst black summer picked
swim	0.78	takes name fish
round	0.78	annual
wings	0.76	wings bird
warm	0.75	heavy
made_of_wood	0.69	wooden
female (control)	0.84	actress birth herself

Table 9.2: Contexts with the top distinctiveness score for each property in the giga corpus. Multiple contexts can have the same score. The mean f1 score represents the mean of the f1-scores resulting from all semantic categories.

9.2.3 Evidence annotation

The evidence candidates extracted on the basis of salience and distinctiveness are particularly likely to impact embedding representations and diagnostic classification results. Furthermore, the extracted candidates are characteristic of the contexts of positive example concepts. They may yield insights into what type of property evidence embedding models could have picked up. They also serve as a dataset for testing theoretical assumptions about the expression of property evidence introduced in Chapter 3.

As a basis for further analysis, I manually categorize the evidence candidates according to types of property-evidence. The categories used for this annotation process are based on the evidence types introduced in Chapter 3 and are summarized again in Table 9.3. For example, property-evidence consist of words directly expressing the property (e.g. **red**: *red*), words that strongly imply the property (e.g. **fly**: *land*), words that are instances of the property (e.g. **blue**: *sea*), or words that are thematically related to the property (e.g. **swim**: *sea*).

specificity	label	explanation	examples
	p	direct expression of the property	**warm**-*warm*, **lay_-eggs**-*eggs*
prop. specific	n	near synonym of the property	**hot**–*heated*
	l	logical or highly likely implication	**hot**-*burning*
	i	instance of the property (i.e. Something that has the property)	**red**–*blood*
non-specific	r	thematically related to the property	**swim**–*sea*
	b	associated with the property via cultural biases	**female**–*beautiful*
unrelated	u	unrelated to the property	**blue**–*magic*

Table 9.3: Types of property-evidence.

	giga							wiki						
	p	n	l	i	r	b	u	p	n	l	i	r	b	u
square	0	0	0	62	2	0	1440	0	0	0	28	0	0	413
warm	4	2	3	90	22	0	1987	1	0	6	61	15	4	1094
black	1	0	1	46	9	0	1092	1	1	2	53	6	0	756
red	1	0	0	89	6	0	1671	1	0	0	91	2	0	1181
fly	6	1	2	8	31	3	901	4	0	1	9	9	0	32
dangerous	2	7	14	44	60	6	981	1	1	4	33	36	0	552
wings	2	0	5	27	38	15	473	0	0	6	28	22	1	275
sweet	2	0	0	19	0	0	14	1	0	0	10	0	0	15
hot	2	1	0	15	7	0	59	2	2	4	11	6	0	67
used_in_cooking	4	2	2	188	173	0	203	2	5	2	132	126	1	437
juicy	2	0	0	88	12	0	708	1	0	0	20	3	0	98
green	1	0	0	87	8	0	482	1	0	0	68	8	0	473
made_of_wood	4	0	0	19	27	0	736	9	0	0	15	19	0	454
blue	2	0	0	157	10	0	2065	1	0	0	136	8	0	1673
yellow	1	0	0	3	0	0	48	0	0	0	8	0	0	103
roll	5	0	5	95	44	0	3737	3	0	7	79	41	0	3355
cold	1	5	0	48	19	0	1008	1	1	0	24	3	0	561
round	1	0	1	38	2	0	478	2	1	1	42	1	0	987
wheels	2	0	2	13	23	0	204	3	0	2	9	18	0	73
lay_eggs	2	0	0	13	7	0	52	0	0	0	9	4	0	13
swim	1	0	0	12	41	0	1327	2	0	0	33	42	0	1623
female (control)	0	0	3	7	0	7	28	1	0	2	14	0	5	87

Table 9.4: Overview of annotation.

The types of evidence fall into two broad categories: property-specific evidence and non-specific evidence. Property-specific evidence indicates a property directly, while non-specific evidence is less targeted and only points to the property indirectly, via lexical associations (e.g. via related concepts such as *pilot* for the property **fly**). This type of evidence can point towards fine-grained semantic categories (e.g. flying vehicles) that largely overlap with instances of the property. Overall, it is a less reliable source of evidence. In addition to property-specific and non-specific evidence, the extraction method also yielded words that are entirely unrelated to the property. Such words can still serve as evidence for a diagnostic classifier. However, it is important to note that they are semantically not related to the property in question. If a diagnostic classifier relied on such expressions, it did not identify evidence of the semantic property.

To ensure that the categorization was done consistently, I applied the following checks:

- The annotations were done separately for each corpus. This means that the same context may have been annotated twice. I checked whether the annotations of the same contexts are consistent between the two corpora.
- Instances of a property should be annotated as a type of property evidence. Positive examples themselves can appear as property instances in the contexts of other positive examples. I used the positive examples from the property datasets to check if all positive examples have been annotated correctly as property instances if they appeared as context words.
- I used manually compiled lists of words expressing the property directly (e.g. **red**: *red*) to identify direct property expressions.

The distribution of evidence types over all evidence candidates is presented in Table 9.4. The largest group of evidence words tends to consist of unrelated words. Property-instances and related words also constitute substantial components. The categorized context words allow for various analyses of the expression of property evidence. The following section introduces different measures of evidence representation in the corpora.

9.2.4 Measuring evidence

In this section, I introduce four different measures of property evidence that can be used to analyze the annotated set extracted contexts presented in the previous section. The goal of the measures is to establish how strongly different types of property evidence are represented in the corpora. Evidence that is represented strongly is more likely to impact embedding vectors and to be identified by diagnostic classifiers than evidence that is represented weakly. Being able to compare property evidence also enables testing specific hypotheses about the expression of property evidence.

Proportion of evidence The extraction method based on salience and distinctiveness yielded sets of evidence candidates per property. For examples, for the property **green**, 578 candidates were extracted from the giga corpus and 550 from the wiki corpus. The candidates were then

annotated with respect to different types of evidence. For instance, the context word *green* was labeled as property-specific evidence, whereas the word *peas* was labeled as property-instance. Contexts entirely unrelated to the property were annotated as unrelated evidence (e.g. *mild*). Based on these annotations, it is possible to calculate the proportion of evidence candidates of a particular evidence type. For instance, the proportion of property-specific evidence can give a first indication of how well property-specific evidence is represented in the corpus data. A high proportion of property-specific evidence words means that many of the salient and distinctive contexts are, in-fact, reflections of the target property. The proportion of property-evidence can be calculated on the level a property (e.g. the entire dataset for **green**) as well as on the level of a specific positive example concept (e.g. *grass*). To calculate proportion of evidence on the level of an individual concept, I count how many of the evidence candidates in the contexts of a particular concept serve as property evidence. This calculation can be performed for each type of property evidence.

Evidence coherence Whether property-evidence can be encoded in embedding vectors is likely to depend on the lexical coherence of evidence words. If all evidence words of a property are semantically very similar, they are more likely to impact the embedding representation of the positive property examples in such a way that the evidence can be recognized by a diagnostic classifier. The most extreme version of lexical coherence would be a single evidence word (e.g. **red**: *red*) that systematically occurs in all contexts of positive examples. If embedding models can indeed abstract over similar evidence words (e.g. **dangerous**: *risky*, *threatening*, *dangerous*), multiple semantically similar evidence words should also be effective. In contrast, semantically different evidence words are unlikely to be encoded systematically in such a way that they are recognizable by a diagnostic classifier. For example, if property-evidence of a color property is primarily expressed by instances of the example contexts (e.g. **blue**: *sea*, *paint*, *car*, *jeans*), it is hardly possible for a diagnostic classifier to detect commonalities between the positive examples. I measure coherence in terms of mean cosine similarity of all possible pairs of evidence words.

Evidence distinctiveness I measure the distinctiveness of individual evidence words by means of their ability to distinguish positive from negative examples. This can be quantified in their precision and recall (summarized as f1 score). The more positive examples an evidence word can distinguish from the negative examples of a property, the higher the chances that enough vectors of positive examples carry property information that could be identified by a diagnostic classifier. Each of the extracted contexts has a distinctiveness score. The distinctiveness of a set of contexts (e.g. all property-specific evidence words of the property **fly**) can be analyze by calculating the mean distinctiveness score over all property-specific evidence words.

Evidence strength I establish how strongly property evidence is represented in the context of a concept by means of its raw cf-iwf score. If an evidence word is expressed frequently in the context of a concept (and infrequently in the contexts of the negative examples) it receives a high cf-iwf score. The raw scores can be used on the level of individual concepts.

To analyze a set of examples, the strength scores of individual examples can be summarized by the mean score.

The four measures presented in this section are by no means independent of one another. Rather, they are likely to be correlated (e.g. a high cf-iwf score is likely to be correlated with high distinctiveness). However, they can highlight slightly different characteristics of the property evidence.

9.3 Analysis 1: Property-evidence and diagnostic classification

In this section, I explore the relation between property-evidence and successful diagnostic classification. The context words extracted based on their salience and distinctiveness are likely to impact embedding representations. However, not all of them can be expected to impact the embeddings equally strongly. The goal of this section is to gain deeper insights into what type of property evidence is most likely to have been detected by diagnostic methods (discussed in Chapter 8). If the diagnostic classifiers have detected evidence of the semantic property in question (e.g. **fly**), property-specific evidence (*fly*, *flew*, *land*) should be represented in the underlying corpus data most strongly. If the diagnostic classifiers have detected information about fine-grained semantic categories, property-instances (e.g. *robin*, *sparrow*, *airplane*) and related words (e.g. *nest*, *pilot*) should be represented strongest. If the classifiers have relied on accidental correlations rather than information about the property, words unrelated to the property should be represented best. Before presenting the results of this comparison, I describe the details of the comparison setup.

9.3.1 Measuring Evidence Representation on the Level of Properties

The goal of this analysis is to establish what type of property evidence is likely to have an impact on embedding representations and to be picked up by diagnostic classifiers. To analyze property evidence candidates, I calculate the measures presented in Section 9.2.4 on the level of individual properties. This enables a comparison between different property types. In particular, I distinguish between property-specific evidence (direct property expressions, near synonyms and logical or highly likely implications), non-specific evidence (property instances, related words, and social biases), and unrelated words (refer to Section 9.2.3 for a detailed overview of evidence types). The measures are calculated as follows:

- Evidence proportion: The proportion of evidence words of a particular type (e.g. property-specific) out of all extracted evidence word candidates (between 0 and 1).
- Coherence: Mean cosine similarity of all possible pairs of evidence words of a particular evidence type (e.g. property-specific) (between 0 and 1).
- Mean and maximum evidence distinctiveness: The mean f1-score of evidence words of a particular type and the score of the best performing evidence word of a particular evidence type (between 0 and 1).

- Mean and maximum evidence strength: The mean cf-iwf score of evidence words of a particular type and cf-iwf score of the evidence word of a particular type with the highest score (between 0 and 1).

The measures enable a comparison between different evidence types. Each score can be calculated for each of the three types of evidence under consideration (property-specific, non-specific, unrelated). The scores are only meaningful when put into comparison to one another. It is unclear what level of evidence proportion, coherence, and strength of evidence words is required for them to impact embedding representations and to be detected by a diagnostic classifier.

To illustrate the behavior of the different measures, I show the results for nine properties in the giga corpus in Table 9.5. The table shows the scores for property-specific evidence per property. The results differ between the properties: the control property **female** clearly 'wins' in terms of proportion, but does not score highest for any of the other measures. In contrast, **juicy** and **lay_eggs** have a lower proportion, but score comparatively highly for the other scores. In general, it can be observed that the scores vary; scoring highly for one measure does not necessarily mean scoring highly for all measures. This supports the intuition that the scores highlight slightly different aspects.

	proportion	coherence	dist-mean	dist-max	str-mean	str-max
dangerous	0.0206	0.2389	0.7224	0.8404	0.0034	0.0095
swim	0.0007	**1.0000**	0.6726	0.6726	0.0035	0.0035
fly	0.0095	0.3466	**0.8334**	0.9007	0.0043	0.0084
black	0.0017	0.1526	0.7047	0.7504	0.0079	0.0141
used_in_cooking	0.0140	0.3850	0.8230	**0.9308**	0.0103	0.0216
lay_eggs	0.0270	0.8302	0.8303	0.8432	0.0113	0.0133
wheels	0.0164	0.3856	0.7924	0.8432	0.0117	0.0212
juicy	0.0025	0.6378	0.7923	0.7927	**0.0339**	**0.0616**
female (control)	**0.0667**	0.5798	0.6333	0.6949	0.0049	0.0078

Table 9.5: Results of the evidence metrics for four properties in the giga corpus (raw numbers).

The scores of a property are only informative when compared to other properties. To facilitate the interpretation of the different metrics, I represent the results in terms of distance to the median score calculated over all properties. Scores above the property-median will thus receive a positive value and scores below the median a negative value. To make the scores comparable, I normalize the distance by representing it as the proportion of the median. For instance, if the median proportion score is 0.3 a property with a raw cosine score of 0.3 will have a distance of 0. A property with a cosine score of 0.6 will have a distance of 1, a property with a diversity score of 0.4 will have a distance of 0.33 and a property with a cosine score of 0.1 will receive a distance of -0.66.

The normalized scores can be summed to show a total evidence representation score. I sum over all measures and calculate the mean. To soften the impact of outliers, I also calculate a binary overall score, which simply counts how many measures lie above the median.

	prop.	coh.	dist.		str.		sum	bin
			mean	max	mean	max		
sweet	9.74	0.21	0.1	0.12	0.17	0.18	1.75	1.00
hot	5.71	-0.28	-0.03	-0.02	1.55	0.83	1.3	0.50
juicy	-0.54	0.65	0.06	0	3.3	3.38	1.14	0.67
lay_eggs	4.08	1.15	0.11	0.06	0.43	-0.06	0.96	0.83
yellow	2.61	1.59	0.06	-0	0.91	0.07	0.87	0.83
red	-0.89	1.59	0.06	0	2.09	0.73	0.6	0.83
wheels	2.08	0	0.06	0.06	0.48	0.51	0.53	0.83
used_in_cook-ing	1.63	-0	0.1	0.17	0.31	0.53	0.46	0.83
green	-0.67	1.59	0.15	0.08	0.91	0.07	0.36	0.83
made_of_-wood	-0.04	0.14	0	0.02	0.79	0.73	0.27	0.67
dangerous	2.88	-0.38	-0.03	0.06	-0.57	-0.33	0.27	0.33
wings	1.35	0.02	0.02	0.08	-0.27	-0.41	0.13	0.67
fly	0.78	-0.1	0.12	0.14	-0.46	-0.41	0.01	0.50
cold	0.04	-0.13	-0.11	-0.02	-0.14	-0.2	-0.09	0.17
swim	-0.86	1.59	-0.1	-0.15	-0.56	-0.75	-0.14	0.17
blue	-0.83	-0.24	-0.05	-0.03	-0.01	-0.1	-0.21	0.00
roll	-0.52	-0.46	-0.1	-0.01	-0.34	0.11	-0.22	0.17
black	-0.67	-0.6	-0.06	-0.05	0	0	-0.23	0.00
warm	-0.2	-0.3	-0.23	-0.06	-0.5	-0.4	-0.28	0.00
round	-0.28	-0.68	-0.21	-0.14	-0.4	-0.41	-0.35	0.00
square	-	-	-	-	-	-	-	0.00
female (con-trol)	11.53	0.5	-0.15	-0.12	-0.38	-0.44	1.82	0.33
med. (raw)	0.01	0.39	0.75	0.79	0.01	0.01	-	-

Table 9.6: Results of the evidence metrics in the giga corpus (normalized distance scores) for property-specific evidence.

Both summed scores attribute the same weight to all metrics, which is not necessarily a fair comparison, as some factors could have a higher impact on the embedding representations than others. Nevertheless, the summed scores can provide a first indication of property-evidence representation in the corpus data.

Table 9.6 represents the normalized scores for all properties in the giga corpus. The final two columns show the summed scores. The properties in the table are ranked by their overall summed score. According to the summed score, the control property **female** has the highest representation of property evidence, while **round** has the lowest. Despite the overall strong score of **female**, it scores below the median for evidence distinctiveness and strength. When considering the binary overall score (which punishes low performance on individual measures) it can be observed that **sweet** scores highest, followed by **lay_eggs**, **used_in_cooking**, **wheels**, **yellow**, and **green**. There is no property-specific evidence for **square**.

To facilitate this analysis, I interpret the evidence representation scores as follows: If the summed score is negative or the binary summed score is below 0.5 (i.e. less than half of the

	prop-specific-sum	prop-specific-bin	non-specific-sum	non-specific-bin	u-sum	u-bin	cl	evidence type
warm	-0.28	0.00	-0.08	0.17	0.11	0.33	False	
black	-0.23	0.00	-0.30	0.00	0.02	0.33	False	
red	0.6	0.83	-0.05	0.17	-0.06	0.50	False	prop-specific
dangerous	0.27	0.33	-0.06	0.50	0.00	0.50	False	
round	-0.35	0.00	0.09	0.50	0.07	0.50	False	
cold	-0.09	0.17	-0.08	0.17	0.23	0.67	False	u
hot	1.3	0.50	0.68	0.67	0.06	0.33	False	non-specific
roll	-0.22	0.17	-0.22	0.00	0.05	0.67	False	u
green	0.36	0.83	0.26	0.67	-0.06	0.33	False	non-specific prop-specific
made_of_wood	0.27	0.67	-0.20	0.00	-0.16	0.17	False	prop-specific
blue	-0.21	0.00	-0.21	0.17	0.01	0.67	False	u
yellow	0.87	0.83	-0.01	0.33	-0.05	0.67	False	prop-specific
square	-	0.00	0.25	0.83	0.29	1.00	True	non-specific u
wheels	0.53	0.83	0.69	1.00	0.08	0.67	True	non-specific prop-specific u
female	1.82	0.33	0.92	1.00	0.06	0.50	True	non-specific
juicy	1.14	0.67	0.42	1.00	0.15	0.50	True	non-specific prop-specific
used_in_cooking	0.46	0.83	1.65	1.00	-0.12	0.50	True	non-specific prop-specific
sweet	1.75	1.00	1.37	1.00	-0.25	0.33	True	non-specific prop-specific
wings	0.13	0.67	-0.01	0.33	-0.10	0.17	True	prop-specific
fly	0.01	0.50	-0.29	0.33	0.02	0.67	True	u
lay_eggs	0.96	0.83	0.57	0.83	0.05	0.50	True	non-specific prop-specific
swim	-0.14	0.17	-0.10	0.33	-0.04	0.50	True	

Table 9.7: Comparison to diagnostic classification (giga).

	prop-specific-sum	prop-specific-bin	non-specific-sum	non-specific-bin	u-sum	u-bin	cl	evidence type
warm	-0.32	0.17	-0.07	0.33	0.24	0.50	False	
black	-0.07	0.17	-0.29	0.00	0.14	0.33	False	
red	0.37	0.67	-0.16	0.17	-0.08	0.33	False	prop-specific
round	-0.3	0.00	-0.03	0.50	0.12	0.50	False	
cold	-0.24	0.00	-0.24	0.17	0.17	0.67	False	u
sweet	1.47	1.00	1.05	0.83	-0.19	0.33	False	non-specific prop-specific
hot	2.77	0.67	0.42	1.00	0.03	0.33	False	non-specific prop-specific
roll	0.04	0.33	-0.19	0.17	0.17	0.67	False	u
green	0.21	0.67	-0.03	0.17	-0.08	0.17	False	prop-specific
blue	0.26	0.67	-0.15	0.33	-0.03	0.50	False	prop-specific
square	-	0.00	0.28	0.83	0.01	0.83	True	non-specific u
wheels	1.6	0.83	0.84	1.00	0.29	0.83	True	non-specific prop-specific u
female	0.57	0.33	0.60	1.00	0.03	0.67	True	non-specific u
yellow	-	0.00	-0.15	0.33	-0.04	0.50	True	
juicy	1.2	0.83	0.55	0.83	0.12	0.50	True	non-specific prop-specific
lay_eggs	-	0.00	0.89	0.67	0.39	0.50	True	non-specific
used_in_cooking	0.38	0.83	0.77	1.00	0.01	0.67	True	non-specific prop-specific u
wings	0.33	0.67	0.04	0.50	0.04	0.67	True	prop-specific u
dangerous	-0.05	0.17	-0.00	0.33	0.10	0.50	True	
fly	2.49	0.50	0.41	0.67	-0.01	0.33	True	non-specific
made_of_wood	0.55	0.67	-0.14	0.00	-0.11	0.17	True	prop-specific
swim	-0.3	0.17	-0.11	0.17	-0.01	0.50	True	

Table 9.8: Comparison to diagnostic classification (wiki).

measures score below the median), I interpret the property-representation as low. In all other cases, I count property-representation as high. It should be noted that this cut-off does not necessarily indicate sufficient evidence for property encoding in the embeddings. Rather, it was chosen as a transparent threshold. I use this threshold to compare the representations of different evidence types to the diagnostic classification outcomes. Other methods of determining 'high' and 'low' property evidence might result in a better fit in such a comparison. In this study, however, I do not aim to find a perfect fit. Rather, I aim to present a global overview of evidence types and their potential impact on embedding representations. Using a simple and transparent cut-off reduces the risk of 'cherry-picking' results. If the diagnostic classifiers could identify property-specific information, this should arise from the comparison.

To illustrate the comparison of property-evidence, I show the results of the aggregated scores for all properties and evidence types in Table 9.6 for the giga corpus. The results show that properties differ considerably with respect to how strongly they are represented by property-specific evidence in the giga corpus. The scores show that property-representation in corpus data does not necessarily depend on property categories; for instance, not all color-properties have low scores (which is expected on the basis of the hypotheses introduced in Chapter 3). The two shape-properties, however, are indeed hardly represented.

9.3.2 Comparison to Classification Performance

In this section, I compare the different types of linguistic property evidence found in the corpora to the outcome of the diagnostic classification experiments (presented in Chapter 8). If the diagnostic classifiers could identify property-specific information in the embeddings, then the evidence found in corpus data for this property should be particularly well represented. To measure this, I use evidence proportion, coherence, distinctiveness, and evidence strength as illustrated in the previous section. Successful diagnostic classification is defined as clearly having outperformed the control task baseline (see Chapter 8 for details about the control task).

The analysis presented in this section also acts as a 'sanity check' for the evidence measures; properties with successful diagnostic classification results have to score highly for at least one of the evidence types. If they do not, the measures to not capture relevant factors that impact embeddings. It should be noted that a perfect fit (i.e. high evidence representation for all successfully classified properties) cannot be expected due to the definition of high evidence representation on the basis of the median scores. The general tendency, however, should hold.

Aggregated results

Table 9.7 shows the results of the evidence analysis compared to the results of the diagnostic classification experiments in the giga corpus and Table 9.8 shows the results in wiki. The tables show the normalized summed and binary scores for the evidence types 'property-specific', 'non-specific', and 'unrelated' for properties classified successfully ('True') and properties not classified successfully ('False'). The final column summarizes the evidence types that score above the threshold (positive summed score, binary score > 0.5). If the

classifiers have indeed picked up property-specific evidence, then the successfully learned properties should score highly for property-specific evidence. The unsuccessfully classified properties should not score highly for any evidence type.

Giga corpus The following observations can be made about property evidence in the giga corpus (Table 9.7): As expected, for four unsuccessfully classified properties, it can indeed be observed that none of the evidence types are particularly well represented in the corpus data (**warm**, **black**, **dangerous**, **round**). In contrast, for all successfully classified properties except **swim**, at least one evidence type is represented well. The fact that the overall tendency holds is an indication that the metrics used to quantify property evidence can, indeed, reflect how likely information is to be picked up by a diagnostic classifier.

When comparing the strongly represented evidence types of successfully classified properties to the unsuccessfully classified properties, the following differences can be observed: Firstly, for unsuccessfully classified properties, only one evidence type is well represented. In contrast, for six (out of ten) successfully classified properties, more than one evidence type is well represented. Secondly, for six unsuccessfully classified properties, property-specific evidence is the only evidence type that is well represented. When considering the successfully classified properties, it can be observed that property-specific evidence hardly ever appears as the only well-represented evidence type. In the majority of successfully classified properties, non-specific evidence (i.e. property-evidence that is not a direct reflection of the property or logically linked to it) is well represented (eight out of ten). In contrast, non-specific evidence is only well-represented for two out of twelve unsuccessfully classified properties. These tendencies indicate that the diagnostic classifiers are more likely to have picked up information about property instances of thematically related words than hard, property-specific evidence.

The following exceptions to the overall tendencies can be observed: For property **swim** none of the evidence types seem to be represented strongly, despite the successful diagnostic classification result. A possible explanation for this could be that the threshold was set too strictly or that the aggregated scores do not give enough weight to individual measures. An inspection of the individual scores for swim showed that property-specific evidence scores particularly highly (normalized score of 1.59) for coherence. It is possible that highly similar property-specific evidence words together impacted the representations. Interactions between words cannot be reflected by the scores that only consider individual words (all scores except coherence). Individual scores for other evidence types (non-specific and unrelated) also show high scores. It is thus also possible that a combination of evidence words from different evidence types together impacted the embeddings and led to successful classification. This cannot arise from the analysis conducted here.

A a second exception can be observed for the successfully classified property **wings**. Its only strongly represented evidence type is property-specific evidence. It is possible that the property **wings** is indeed a case in which property-specific evidence is strong enough to be represented in embeddings. It should also be kept in mind that different types of evidence can interact. As in the case of **swim**, it could be the case that the combination of property-specific evidence words and other evidence words (e.g. instances of the property) together impacted the embedding representations and were picked up by diagnostic classifiers.

Wiki corpus When considering the results in the wiki corpus (Table 9.8), similar tendencies can be observed. However, it should be noted that three (out of twelve) successfully classified properties do not show any well-represented evidence types (**swim**, **dangerous**, **yellow**). As in the giga corpus, possible reasons for this could the following: The thresholds based on the median were not fitted to the results and may simply be too high (e.g. unrelated evidence is close to the threshold for **dangerous**). In addition,the embeddings could reflect interactions between words of different evidence types.

		proportion	coherence	dist-mean	dist-max	str-mean	str-max
prop-specific	mean-diff	0.0121	0.0570	-0.0175	-0.0393	-0.0012	0.0008
	median-diff	0.0092	0.1438	0.0757	0.0611	-0.0003	-0.0037
	std	0.0233	0.2783	0.2538	0.2740	0.0094	0.0175
non-specific	mean-diff	0.1423	0.0546	0.0973	0.1101	0.0027	0.0107
	median-diff	0.0849	0.0858	0.1286	0.1019	0.0036	0.0120
	std	0.2106	0.0707	0.0613	0.0452	0.0037	0.0187
u	mean-diff	-0.1545	0.0073	0.0443	0.0253	0.0003	-0.0041
	median-diff	-0.0881	0.0076	0.0710	-0.0105	0.0003	-0.0085
	std	0.2249	0.0123	0.0454	0.0941	0.0010	0.0165

(a) giga

		proportion	coherence	dist-mean	dist-max	str-mean	str-max
prop-specific	mean-diff	0.0045	-0.1984	-0.1232	-0.1380	-0.0024	-0.0040
	median-diff	0.0075	0.0358	0.0592	0.0026	-0.0045	-0.0069
	std	0.0265	0.2917	0.3412	0.3665	0.0097	0.0120
non-specific	mean-diff	0.0817	0.0411	0.0899	0.0917	0.0015	0.0050
	median-diff	0.0915	0.0417	0.1058	0.1007	0.0013	0.0006
	std	0.1429	0.0539	0.0679	0.0687	0.0040	0.0146
u	mean-diff	-0.0863	0.0125	0.0430	0.0425	0.0011	-0.0091
	median-diff	-0.1083	0.0120	0.0421	0.0442	0.0005	-0.0109
	std	0.1529	0.0221	0.0590	0.0717	0.0020	0.0058

(b) wiki

Table 9.9: Differences between the mean and median scores of successfully classified properties and unsuccessfully classified properties for specific measures in the giga and wiki corpus.

9.3.3 Results per measure

What is clearly visible from both corpora is that property-specific evidence by itself does not seem to provide a strong signal to diagnostic classifiers. For most successfully classified properties, multiple types of property-evidence scored highly. In most cases, the highly scoring evidence types involve non-specific evidence (i.e. property instances and words related to the property or its instances).

The analysis presented so far is based on normalized measure scores and somewhat arbitrary thresholds. To test whether the tendencies observed so far are also reflected by the absolute scores, I compare the mean scores of successfully and unsuccessfully classified properties for different evidence types. If an evidence type impacts the outcome of the diagnostic classification results, it should show high scores for successfully classified properties and low scores for unsuccessfully classified properties. The differences between the scores are presented in Table 9.9 for both corpora. I opt against conducting significance tests on the rather small data. Instead, I show the differences between the mean and median scores.

Giga corpus The following observations can be made for property evidence in the giga corpus (Table 9.9a): Overall, the initially observed trend seems to hold: For non-specific evidence, successfully classified properties score higher than unsuccessfully classified properties across all measures. When considering the measures in detail, it can be observed that property-specific evidence does show the highest differences for coherence (i.e. property-specific evidence words tend to be more coherent than non-specific or unrelated words). This is to be expected, as the group of property-specific evidence words is usually small and closely tied to the property itself, whereas non-specific evidence words can belong to a variety of semantic categories. For all other scores, non-specific evidence clearly shows the highest differences.

Wiki corpus The same general trend can also be observed in the wiki corpus (Table 9.9b); non-specific evidence shows the highest differences between successfully and unsuccessfully classified properties. In contrast to giga, non-specific evidence words also scores highest for coherence. This is surprising, as non-specific evidence is not necessarily semantically coherent.

9.3.4 Discussion

From the analysis conducted in this section, it is not possible to determine whether there is a causal relationship between this type of evidence and successful classification. A possible way of establishing whether there is indeed a causal relationship could be corpus evidence manipulation. At this point, it can be concluded that the evidence the classifiers are most likely to pick up is non-specific evidence consisting of property instances (e.g. **red**: *blood*, *paint*, *car*) and words related to the property (e.g. **fly**: *pilot*, *airport*, *crew*) through thematic relations or social biases. This tendency is in line with the observations made in the error analysis of the diagnostic experiments; the results indicated that classifiers may pick up evidence of fine-grained semantic categories, rather than specific properties. Such categories are likely to be reflected well through hyponyms and hypernyms, which are part of non-specific property evidence.

9.4 Analysis 2: Hypotheses about Property-Specific Evidence

In this section, I analyze the expression of properties with respect to the factors that may impact whether properties are expressed: Property-concept relations, property types, and

factor	relation	example	expressed
impliedness	`implied_category`	**mammal** - *cat*	no
variability and specification	`variability_limited`	**yellow** - *bell pepper*	yes
	`variability_open`	**red** - *car*	no
typicality	`typical_of_property`	**red** - *blood*	yes
	`typical_of_property`	**green** - *broccoli*	no
affordedness	`afforded_usual`	**fly** - *seagull*	yes
	`afforded_unusual`	**swim** - *cat*	no
	`affording_activity`	**round** - *bowling ball*	yes

Table 9.10: Overview of property-concept relations and hypotheses about whether property information is likely to be expressed in corpora.

.

genre. Property-concept relations form the central component of the theoretical model introduced in Chapter 3. The central hypothesis of this model is that whether a property is expressed in texts depends on the relation that holds between the property and a particular concept. For example, the model predicts that properties that are part of the highly implied knowledge about a concept are unlikely to be expressed (e.g. **mammal**: *cat*). This relationship between property and concept is expressed by the relation `implied_category`. Variable properties, in contrast, are expected to be expressed (e.g. **red**: *apple*). The property-concept relations and hypotheses are summarized in Table 9.10 and explained in detail in Chapter 3. In this section, I use the measures defined in Section 9.2.4 to test these hypotheses on the basis of the same extracted and annotated context words used in the previous analysis. For this analysis, I compare the expression of property-specific evidence words between property-concept relations. I only focus on property-specific evidence words, as this type of property evidence is the only 'hard' evidence that specifically points to a property, rather than a semantic category.

In addition to property-concept relations, other factors may also impact what type of information is expressed in corpora. Previous research has argued that whether property information is expressed in corpus data depends on the type of property. For instance, it has been shown that visual information (e.g. about colors and shapes) is less well represented in context-free distributional models (e.g. Rubinstein et al., 2015) (refer to Chapter 1 for more details). In addition to property types, it can be expected that the genre of a corpus impacts what type of conceptual information it will emphasize. For example, encyclopedic texts can be expected to place more emphasis on making conceptual knowledge explicit, while news texts can be expected to focus on events. I also test these expectations using the context words annotated as property-specific evidence.

For this analysis, I only use measures that consider the degree to which evidence words are being expressed. Whether they have an impact on the embedding representations is not the focus of this analysis. Thus, I only consider the following measures:

- the proportion of property-specific evidence among the candidates

- the strength of property-specific evidence

The remainder of this Section is structured as follows: Section 9.4.1 presents an analysis of property-specific evidence for property-concept relations. This is followed by an analysis of specific properties (Section 9.4.2) and genre (Section 9.4.3) with respect to property-specific evidence. The discussion of the results will be taken up in a general discussion at the end of the chapter (Section 9.5).

9.4.1 Property-Concept Relations

The model of property-evidence expression proposed in Chapter 3 is based on linguistic factors depending on relations that hold between properties and concepts. According to this model, whether property evidence is likely to be expressed depends on how the relation between the property and a specific concept. A summary of the factors and the resulting property concept relations is shown in Table 9.10 (for a detailed explanation, refer to Chapter 3). Before presenting the results for specific hypotheses derived from the model, I outline how property-concept pairs were assigned to specific relations and how property-evidence was analyzed on the level of individual relations.

Relation Assignment

The analysis of individual relations is complicated by the fact that multiple relations can be assigned to a single property-concept pair. As shown in Chapter 7, the annotations in the diagnostic dataset resulted in complex interactions of relations for individual property-concept pairs. Therefore, I use a general comparison and two strategies to assign property-concept pairs to relations.

General Comparison At its core, the theoretical framework predicts the following: Corpora should contain property-evidence for a property-concept pair if it has been annotated with any of the relations expected to trigger property evidence. For all other property-concept pairs, corpora should contain less or no evidence. For instance, the pair **red**-*cherry* is annotated with the following relations: `typical_of_concept`, `typical_of_property`, `implied_category`, `variability_limited`. `Typical_of_concept` and `implied_category` are not expected to trigger evidence, but `typical_of_property` and `variability_limited` are. Therefore, it can be expected that the contexts of the word *cherry* contain evidence of the property **red**. In contrast, the pair **red**-*couch* is annotated with the relation `variability_open`. The relation is not expected to trigger evidence. Hence, it can be expected the contexts of the word *couch* contain less evidence of the property **red**. This type of comparison does not require assigning specific relations. Instead, pairs are simply assigned to one of two possible expectations. This has the disadvantage that the analysis does not provide insights about specific relations.

Strict Relation Assignment To provide a controlled analysis of relations, I apply a strict approach to the selection of pairs for a specific relation: In this approach, I only select pairs for

which the relation under consideration is the only relation that describes the property-concept pair. This has the advantage of eliminating possible interactions. It has the disadvantage that such cases comparatively infrequent in the dataset. For example, the relation `implied_category` only appears in 16 pairs as the only relation assigned to a pair. Insights based on the 'strict' relation assignment method thus tend to be based on limited data.

Loose relation assignment To exploit a larger set of examples per relation, I use a second strategy for assigning property-concept pairs to specific relations. This second strategy uses looser requirements than the strict assignment method. For relations expected to trigger property-evidence, I select all pairs for which the relation under consideration is the only relation expected to trigger property-evidence. For example, the pair **red**-*brick* is annotated with the following relations: `typical_of_concept`, `variability_limited`, and `variability_open`. The only relation expected to trigger property evidence is `variability_limited`. Hence, applying loose relation assignment, the pair **red**-*brick* is assigned to the relation `variability_limited`. Pairs in which multiple relations can trigger property evidence (e.g. the example of **red**-*cherry* above) are excluded from the analysis.

What is more challenging is the selection of pairs for relations not expected to trigger property-evidence. A single pair can be annotated with multiple relations that are not expected to trigger property evidence. For example, the pair **blue**-*paint* is annotated with `variability_open` and `typical_of_concept`. Neither of the two relations are expected to trigger property evidence. To assign the pair to a specific relation, I chose the relation that received the highest number of votes in the annotation task (in this case `variability_open`). Pairs for which multiple relations have the same number of votes are excluded. This 'loose' approach has the disadvantage that possible interactions between relations not captured by the initial framework may distort the results.

Evidence Analysis on the Level of Relations

The hypotheses derived from the model of conceptual knowledge and property expression rely on the comparison of property evidence between specific property-concept relations. To test these hypotheses, I aggregate property-evidence on the level of property-concept relations (evidence proportion and evidence strength). I aggregate both scores by calculating the median score over all pairs assigned to a relation. I opt for the median instead of the mean because it is less strongly influenced by outliers (in particular, when dealing with few data points). In the remainder of this section, I present the results of the analysis per hypothesis. For each hypothesis, the most important assumptions are summarized briefly. A detailed overview of all hypotheses derived from the model of property evidence can be found in Chapter 3.

General comparison As discussed in the previous section, assigning property-concept pairs to specific relations is difficult, as individual pairs have been annotated with multiple relations. However, it is possible to test the following, general hypothesis without assigning pairs to relations: Pairs annotated with any of the relations expected to trigger property

	prop.	str-median	str-max	n
evidence_pos	0.0075	0.0074	0.0155	798
no_evidence_pos	0.0043	0.0039	0.0087	470

Table 9.11: Evidence representation on the level of relations in the giga corpus using a general comparison.

	prop.	str-median	str-max	n
evidence_pos	0.0057	0.0063	0.0140	894
no_evidence_pos	0.0026	0.0042	0.0089	502

Table 9.12: Evidence representation on the level of relations in the wiki corpus using a general comparison.

evidence should have more property evidence in the corpora than the pairs not annotated with relations expected to trigger property evidence. The results of this comparison are shown in Table 9.11 for the giga corpus and in Table 9.12 for the wiki corpus. The tables show the results of evidence proportion and evidence strength for property-concept pairs expected to trigger evidence ('evidence_pos') compared to property-concept pairs also annotated with positive relations, but not expected to trigger evidence ('no_evidence_pos'). The scores reflect evidence proportion (median) and evidence strength (maximum and median). In addition, the number of property concept pairs for each condition is shown ('n'). It can be observed that the expected tendencies hold in both corpora; pairs expected to trigger evidence score higher for all three measures in both corpora. This can be seen as a first indication that the general hypothesis holds. However, this analysis cannot provide insights about individual property-concept relations.

Results for specific relations The results of the analysis of property-concept relations are presented in Table 9.13 for giga and Table 9.14 for wiki. The tables show the scores for evidence proportion (median) and evidence strength (max and median) in the strict and loose relation assignment condition. In addition, the number of property concept pairs for each condition is shown ('n'). Tendencies that hold consistently should be reflected by all scores and hold in both conditions. A summary of the comparisons is shown in Table 9.15. I discuss the results for each hypothesis in the remainder of this section.

Variability and specification One factor that is expected to lead to expressions of property-specific information is variability among a limited range of properties. For instance, apples tend to be either red, green, or yellow. In such cases, the knowledge about the particular property is not strongly implied. Rather, it could function as a distinguishing feature when, for instance, identifying a particular instance of an apple. It is expected that the relation `variability_limited` triggers more property evidence than the relation `implied_category`. This tendency is confirmed in both corpora in the strict relation assignment condition and partially confirmed in both corpora in the loose relation assignment condition.

	strict				loose			
	prop.	str-median	str-max	n	prop.	str-median	str-max	n
implied_category	0.0029	0.0013	0.0022	14	0.0085	0.0039	0.0082	85
typical_of_concept	0.0060	0.0190	0.0311	3	0.0070	0.0149	0.0286	91
typical_of_property	-	-	-	-	0.0192	0.0089	0.0271	180
affording_activity	0.0128	0.0071	0.0103	4	0.0176	0.0071	0.0168	119
afforded_usual	0.0714	0.0552	0.0721	2	0.0057	0.0021	0.0032	31
afforded_unusual	0.0013	0.0013	0.0017	19	0.0012	0.0013	0.0017	41
variability_limited	0.0030	0.0048	0.0087	104	0.0026	0.0054	0.0092	451
variability_open	0.0042	0.0038	0.0082	171	0.0037	0.0043	0.0090	321

Table 9.13: Evidence representation on the level of relations in the giga corpus.

	strict				loose			
	prop.	str-median	str-max	n	prop.	str-median	str-max	n
implied_category	0.0022	0.0037	0.0043	16	0.0045	0.0035	0.0085	106
typical_of_concept	0.0353	0.0042	0.0051	4	0.0062	0.0096	0.0193	94
typical_of_property	-	-	-	-	0.0152	0.0050	0.0131	215
affording_activity	0.0453	0.0091	0.0097	4	0.0190	0.0081	0.0179	130
afforded_usual	0.0000	0.0000	0.0000	3	0.0000	0.0012	0.0035	54
afforded_unusual	0.0010	0.0012	0.0018	20	0.0014	0.0016	0.0027	42
variability_limited	0.0032	0.0053	0.0097	112	0.0035	0.0065	0.0124	485
variability_open	0.0017	0.0037	0.0069	172	0.0026	0.0046	0.0082	323

Table 9.14: Evidence representation on the level of relations in the wiki corpus.

group	hypothesis	giga		wiki	
		strict	loose	strict	loose
impliedness	implied_category <typical_of_property	n.a.	✓	n.a.	✓
	implied_category <afforded_usual	✓	✗	✗	✗
	implied_category <affording_activity	✓	✓	✓	✓
	implied_category <variability_limited	✓	✓*	✓	✓*
	implied_category <typical_of_concept	✓	✓*	✓	✓
	implied_category <afforded_unusual	✗	✗	✗	✗
	implied_category <variability_open	✓	✓*	✗	✗
property-illustration	typical_of_property >typical_of_concept	n.a.	✗	n.a.	✗
affordedness	affording_activity >afforded_unusual	✓	✓	✓	✓
	afforded_usual >afforded_unusual	✓	✓	✗	✗
variability	variability_limited >variability_open	✓*	✓*	✓	✓

Table 9.15: Summary of hypotheses. ✓indicates that a tendency holds; ✗indicates that a tendency does not hold. * indicates that a tendency is reflected by the majority of scores but not by all scores.

A second hypothesis about variability is related to the range of possible options. While apples tend to have one out of a limited range of colors, t-shirts or cars can have any color. The range of options is virtually unlimited (`variability_open`). While this type of variability is also expected to trigger explicit property mentions, they are most likely less systematic than for `variability_limited`. In other words, it is more likely that corpora contain information about the fact that apples can be red than about the fact that t-shirts can be red. This hypothesis is confirmed in the giga and partially in the wiki corpus in both relation assignment conditions.

Finally, it can be expected that `variability_open`, albeit expected to trigger less consistend evidence than `variability_limited`, still leads to more explicit property mentions than `implied_category`. This expectation is partially confirmed in giga in the strict (but not the loose) condition. It is not confirmed in wiki.

Property-illustration and typicality The relation-framework contains two different relations that express slightly different notions of typicality: Properties can be particularly closely associated with concepts. In other words, some properties are typical of a concept and immediately come to mind when thinking of a concept (e.g. **green** and *broccoli*). Such an association is referred to as `typical_of_concept`. Typicality can also create associations from property to concept: Some concepts are typical examples of a property. Such concepts tend to illustrate the property particularly well (e.g. *blood* is closely associated with the color **red** and can be used to illustrate it). This notion is captured by the relation `typical_of_property`. It should be kept in mind that the crowd had difficulties with distinguishing the two relations. It is thus possible that the annotations do not accurately reflect the two notions. The annotations accurately reflect close associations, but they do not reflect the fine-grained difference between the two relations.

The association captured by the relation `typical_of_concept` is not expected to

trigger systematic evidence. The notion captured by the relation typical_of_property is expected to trigger systematic evidence. Whereas typical_of_concept can occur without its counter-part, this is highly unlikely for typical_of_property. Therefore, the strict selection mode does not allow for testing the hypothesis. In the loose mode, it could not be confirmed in either of the corpora. The relation typical_of_property is expected to trigger more evidence than implied_category. This expectation could be confirmed in both corpora in the loose condition. Typical_of_concept also triggered more evidence in both corpora. These results could indicate that close associations lead to more property expressions than high impliedness. It is, however, not possible to draw conclusions about the fine-grained differences between the two typicality relations as the annotations are not reliable.

Affordedness Properties that express possible and usually performed activities (afforded_usual: **fly**: *seagull*) or enable such activities (affording_activity: **round**-*bowling ball*) are expected to trigger systematic property evidence. Properties expressing possible but not usually performed activities are not expected to trigger systematic evidence (e.g. **swim**-*cat*). It should be noted that the crowd judgments for afforded_unusual showed a comparatively high number of unreliable judgments. Therefore, the comparisons involving the relation afforded_unusual could be distorted.

Based on the reliable data, the following observations can be made: Properties that afford activities clearly yield more evidence than highly implied information in both corpora and in both conditions. The relation afforded_usual, however, only triggers more evidence in the giga corpus under the strict condition. This is not confirmed in the wiki corpus.

Based on the imperfect data, the following observations can be made: In wiki, only affording_activity shows more evidence than afforded_unusual (in both conditions). Afforded_usual triggers more evidence than afforded_unusual in the giga corpus (in both conditions), but not in wiki. Afforded_unusual does not trigger more evidence than highly implied information.

Summary The analysis of specific property-concept relations yielded the following insights: Overall, relations expected to trigger property evidence do indeed seem to do so in an overall comparison. When considering individual hypotheses, the following tendencies could be observed: Firstly, the results indicate that variability, in particular variability among a limited range of property-options triggers systematic property-evidence. Secondly, properties that afford activities also seem to be mentioned explicitly. Thirdly, close associations between properties and concepts seem to trigger property-evidence. Unfortunately, however, it is not possible to distinguish the two directions of typicality involved in such close associations. Other hypotheses, however, could not be confirmed. Partially, this could be due to unreliable annotations (in particular for the afforded_unusual).

9.4.2 Properties

Previous work on semantic properties in distributional representations has shown that certain properties tend to be present in distributional representations while others tend to be absent.

Rubinstein et al. (2015) showed that perceptual information tends to be absent, while taxonomic information tends to be represented well. The experiments presented in the previous chapter (Chapter 8) are in line with this observation. However, the experiments have also cast doubt on whether the properties themselves are represented. A likely alternative explanation is that distributional representations carry information about fine-grained semantic categories that happen to correlate with semantic properties. The analysis presented in Section 9.3 has shown that diagnostic classifiers are indeed more likely to have picked up information about fine-grained semantic categories than about specific properties. This could also explain observations made in previous research. In this section, I examine to what degree the corpora contain property-specific evidence for particular properties.

For this analysis, I examine property-specific evidence in terms of evidence proportion and evidence strength (summarized in Table 9.16). The proportion of evidence words shows how many of the extracted evidence candidates could be identified as property-specific evidence. The evidence strength is expressed by the raw cfiwf scores (mean and maximum). A high evidence proportion indicates that the context comparison did indeed yield relevant evidence candidates. Evidence strength provides an indication of frequency for positive examples compared to negative examples. Properties that are represented well should score highly for all measures. In addition, I show the absolute number of words expressing the property ('abs.'). The top three scores per measure are shown in bold font.

The scores for individual properties partly confirm the initial expectation. Most color and shape properties score lower than other properties (in particular the part properties and the complex properties), and the taxonomic property **lay_eggs** (in giga)). However, this pattern is not consistent; the color property **red** has one of the top scores for evidence strength, the color property **yellow** is also represented comparatively strongly in the giga corpus, and the taxonomic property **lay_eggs** is not represented at all in the wiki corpus. Other perceptual properties are also represented comparatively well (e.g. **hot** and both taste properties in both corpora). Overall, it is difficult to identify clear patterns for specific types of properties. Whether property-specific evidence is expressed in corpus data does not seem to depend on the property type alone.

It is possible that the property-concept relations present a better explanation of whether property-specific evidence is expressed in corpora. For example, it is likely that the property-concept relation `variability_limited` triggered many of the explicit property mentions for the property **red**. Likewise, it is likely that the property **sweet** is relevant for activities and functions. At this point, however, it is difficult to determine whether the differences observed between different property-concept relations are indeed more meaningful than the differences observed between the property types.

9.4.3 Genres

In this section, I consider differences between the two genres represented by giga and wiki based on the analyses presented above. Chapter 3 outlined hypothesized differences between the two genres: Implied information is expected to be represented more strongly in the encyclopedic texts in the wiki corpus than in the news texts in the giga corpus. Information

		giga				wiki			
		prop.	abs.	str-mean	str-max	prop.	abs.	str-mean	str-max
perceptual (color)	black	0.002	2	0.008	0.014	0.005	4	0.007	0.017
	blue	0.001	2	0.008	0.013	0.001	1	0.013	0.013
	green	0.002	1	0.015	0.015	0.002	1.0	0.011	0.011
	red	0.001	1	**0.024**	**0.024**	0.001	1	**0.017**	0.017
	yellow	0.019	1	0.015	0.015	0.000	0	0.000	0.000
perceptual (material)	made_of_wood	0.005	4	0.014	**0.024**	0.018	**9**	0.011	0.023
perceptual (shape)	round	0.004	2	0.005	0.008	0.004	4	0.005	0.008
	square	0.000	0	0.000	0.000	0.000	0.0	0.000	0.000
perceptual (taste)	juicy	0.002	2	**0.034**	**0.062**	0.008	1	**0.035**	**0.035**
	sweet	**0.057**	2	0.009	0.017	0.038	1	**0.015**	0.015
perceptual (temperature)	cold	0.006	6	0.007	0.011	0.003	2	0.008	0.010
	hot	**0.036**	3	**0.020**	**0.026**	**0.087**	**8**	0.013	**0.029**
	warm	0.004	**9**	0.004	0.008	0.006	7	0.003	0.005
complex	dangerous	0.021	**23**	0.003	0.009	0.010	6	0.005	0.008
	used_in_cooking	0.014	8	0.010	0.022	0.013	9	0.008	0.023
function-action	fly	0.009	**9**	0.004	0.008	**0.091**	5	0.004	0.005
	roll	0.003	**10**	0.005	0.016	0.003	**10**	0.009	**0.027**
	swim	0.001	1	0.003	0.003	0.001	2	0.002	0.003
part	wheels	0.016	4	0.012	0.021	**0.048**	5	0.014	0.026
	wings	0.012	7	0.006	0.008	0.018	6	0.005	0.008
taxonomic	lay_eggs	**0.027**	2	0.011	0.013	0.000	0	0.000	0.000

Table 9.16: Property-specific evidence representation on the level of properties in the giga and wiki corpus. The top three scoring properties per score are marked in bold.

		strict				loose			n_pairs
		prop.	str-mean	str-max	n	prop.	str-mean	str-max	n
afforded_unusual	giga	0.0013	0.0013	0.0017	19	0.0012	0.0013	0.0017	41
	wiki	0.0010	0.0012	0.0018	20	0.0014	0.0016	0.0027	42
afforded_usual	giga	0.0714	0.0552	0.0721	2	0.0057	0.0021	0.0032	31
	wiki	0.0000	0.0000	0.0000	3	0.0000	0.0012	0.0035	54
affording_activity	giga	0.0128	0.0071	0.0103	4	0.0177	0.0071	0.0172	120
	wiki	0.0453	0.0091	0.0097	4	0.0192	0.0081	0.0179	131
implied_category	giga	0.0029	0.0013	0.0022	14	0.0085	0.0039	0.0082	85
	wiki	0.0022	0.0037	0.0043	16	0.0045	0.0035	0.0085	106

Table 9.17: Comparison of relations with respect to genre (giga represents newswire texts, wiki represents encyclopedic texts).

relating to actions (i.e. afforded activities or functions and affording properties) are expected to be represented more strongly in news texts.

Table 9.17 shows a comparison of the impliedness and affordance relation in the two corpora. For impliedness, mixed results can be observed: In the strict condition, the evidence is indeed represented more strongly in the wiki corpus (while the proportions of evidence are similarly high). In the loose condition, evidence is represented almost equally strongly in both corpora (with a higher proportion in giga).

With respect to affordances, the following observations can be made: As expected, the giga corpus clearly shows a higher representation of evidence for the relation `afforded_usual` across both conditions. For `affording_activity`, however, wiki seems to contain stronger evidence representation. It could be the case that the encyclopedic texts place strong emphasis on properties that afford activities, leading to higher evidence representation in the wiki corpus. For unusual actions, no clear differences can be observed.

In addition, the following differences between the corpora can be observed on the level of particular properties (based on Table 9.16 discussed in the previous section): The property **dangerous** is expressed by 23 different words in giga, compared to 6 different words in wiki. A possible explanation for this could be the emphasis of news texts on different types of adverse events (e.g. crime, natural disasters, accidents). Other properties that show striking differences across the two corpora are **yellow** and **lay_eggs**. Both properties are strongly represented in giga, but not in wiki. It is possible that the encyclopedic corpus only mentions the highly taxonomic property **lay_eggs** for specific categories (i.e. the category of BIRD) and uses inheritance for all concepts included in the categories. This would mean that **lay_eggs** is not made explicit for, for instance, specific birds (e.g. *seagull*, *robin*). This observation is another indication that taxonomic property information is most likely encoded via evidence that points to fine-grained semantic categories, rather than property-specific information.

9.5 Discussion and Conclusions

In this chapter, I have used the diagnostic dataset to analyze the expression of properties in two corpora underlying two context-free distributional models. The same models have been examined through diagnostic methods (Chapter 8). The purpose of this chapter was twofold: Firstly, the chapter aimed to verify the results obtained from diagnostic classification using corpus analysis. Secondly, the chapter aimed to give deeper insights about the factors that lead to the expression of property-specific evidence in corpus data.

The results of the diagnostic experiments indicated that the diagnostic classifiers may identify fine-grained semantic categories rather than property specific evidence in the embedding representations. The goal of the first analysis (Section 9.3) was to investigate what type of property-evidence is likely to be identified by diagnostic classifiers.

To find context words in a corpus that are particularly characteristic of a concept, I exploited the contrastive nature of the diagnostic dataset. I compared the contexts of positive examples of a property (e.g. **fly**: **seagull**, *airplane*) to its negative examples (e.g. *penguin*, *bus*) and extracted contexts that are characteristic of the positive examples. This contrastive approach resulted in context words which are likely to (1) have been picked up by diagnostic

classifiers and (2) express property evidence. For instance, the top-ranked words retrieved for the control property **female** include the expressions *herself*, *actress*, and *birth*.

To gain insights about what type of property-evidence is strongly represented in corpus data, I annotated the extracted context words in terms of different evidence types (e.g. property-specific evidence, property-instances, thematically related words). I quantified the degree to which different evidence types are represented in corpus data and compared the results to the outcomes of the diagnostic experiments (found in Chapter 8). The results indicated that diagnostic classifiers are more likely to have picked up non-specific evidence (e.g. things that are **sweet**) than property-specific evidence (e.g. words expressing the property **sweet**). This observation provides additional evidence for the tendency observed in the error analysis of the diagnostic classification results: Distributional data are more likely to represent fine-grained semantic categories than property-specific information. The analysis demonstrated the use of corpus analysis as a means of verifying results obtained from diagnostic experiments.

The comparison between corpus evidence and the results of the diagnostic experiment has an important limitation: The analysis presented in this chapter cannot provide insights into how context-free distributional models (and diagnostic classifiers) react to different types and distributions of property evidence. Ultimately, the analysis cannot indicate with certainty that the models have indeed picked up one type of property-evidence and not another. A possible way of addressing this limitation is to simulate different types of evidence using artificial or manipulated corpus data. The behavior of models trained on differently distributed corpus data could give further insights.

A second purpose of this chapter was to test the hypotheses derived from the model of conceptual knowledge and property evidence presented in Chapter 3. The analysis showed a few tendencies in line with the hypotheses derived from the model: Overall, it seems that property-concept relations that are expected to lead to explicit property evidence do indeed show a stronger representation of property evidence in the corpora. Furthermore, it seems that property-information that affords activities leads to more property expressions than information that is highly implied. Information that is variable also seems to trigger more explicit expressions of property evidence compared to implied information. The latter observation is also in line with contemporary research on contextualized language models and corpus data; as mentioned in Chapter 1, Paik et al. (2021) observe similar tendencies in their exploration of the reporting bias.

Several hypotheses could not be confirmed, in particular in cases for which the annotations of the dataset are not reliable. The crowd annotators could not distinguish close associations from concept to a property (`typical_of_concept`) from close associations from property to concept (`typical_of_property`). The corpus data indicate that close associations do seem to lead to stronger property expressions. However, at this point, it is not possible to determine whether the type of close association plays a role. All results for individual relations are limited to some degree, as they either rely on few data points or run risk of being distorted by complex interaction between relations (see Chapter 7 for details).

In addition to property-concept relations, I also compared property-specific evidence between different properties. Previous research found that perceptual properties tend to be absent from distributional data, while other properties and, in particular, taxonomic properties,

are expressed in distributional data. The analysis of property-specific evidence showed that this pattern holds for some properties, but is by no means consistent. For example, the taxonomic property **lay_eggs** is not mentioned at all in the wiki corpus. Other properties show mixed results. This finding is an additional indication that taxonomic information is encoded via fine-grained semantic category information rather than specific properties. Furthermore, it may be an additional indication that the type of a property does not necessarily determine whether it tends to be expressed in corpora or not.

9.6 Summary

This chapter presented an analysis of property evidence in two corpora. The analysis relied on the contrastive nature of the diagnostic dataset. I extracted candidates of property evidence from contexts of positive examples of a property and annotated them in terms of different types of property evidence (Section 9.2). On the basis of these annotations, I conducted two analyses: Firstly, I compared the property evidence identified in the corpora to the results of the diagnostic experiments presented in Chapter 8 (Section 9.3). The results indicated that diagnostic classifiers are more likely to have picked up fine-grained semantic categories than property-specific information. Secondly, I tested the hypotheses derived from the model of conceptual knowledge (Chapter 3) against the extracted property evidence. The results showed some initial tendencies, but are based on few examples and potentially limited by complex interactions in the data.

10. Challenging Contextualized Language Models

10.1 Introduction

In this chapter, I present an approach towards exploring the semantic knowledge captured by contextualized language models. Rather than using diagnostic classification on vector representations of isolated words, I opt for analyzing the behavior of the models in two challenge tasks. Firstly, I analyze the behavior of pre-trained models using cloze-tasks. Cloze-tasks test the knowledge captured by pre-trained language models by letting them predict a masked token in a sentence. Secondly, I analyze the behavior of models that have been fine-tuned on a Winograd task (using the Winogrande dataset introduced by Sakaguchi et al. (2020))). This task involves complex semantic reasoning. Models fine-tuned on it should be able to reason over semantic properties in Winograd sentences. I present a Winograd-style challenge based on the semantic properties and concepts in the diagnostic dataset (introduced in Part III) and show how it can be used to examine the abilities of fine-tuned models.

If property-knowledge is expressed systematically in corpora, contextualized language models should capture this. A first approach towards testing this is to examine whether pre-trained language models assign different probabilities to positive and negative examples of a property given a context that evokes the property. Consider the following sentence (Example 10.1):

(18) The [mask] is **yellow**.

The sentence evokes the semantic property **yellow** by means of the word *yellow*. A bidirectional language models, such as Bert (Devlin et al., 2019), can use this information to predict candidate tokens for the masked token ([mask]). The model will assign a different probability to each word (or subword) in its repertoire. Thus, it is possible to compare probabilities of different candidates for filling the masked slot (Example 19):

(19) a. The *lemon* is **yellow**.
b. The *sea* is **yellow**.

The task can also be turned around to examine the probabilities assigned to a specific property given the whether a model can predict the property given a concept (Example 20):

(20) a. The *lemon* is [mask].
b. The *sea* is [mask].

A pre-trained language model that has captured information about that fact that lemons tend to be yellow should assign a higher probability to *lemon* than to *sea* in Example 19.

Likewise, it should assign a higher probability to *yellow* in the first sentence of Example 20 than in the second sentence. Analyzing the probability assigned to an individual word is not meaningful, as it is unclear what an individual score would mean. However, the positive and negative examples in the diagnostic dataset allow for a comparative analysis: Overall, positive examples should have a higher probability in a property-evoking context than negative examples. I use this approach to gain first insights into the semantic properties captured by pre-trained contextualized models (Section 10.3). The results indicate that pre-trained Bert models can indeed reflect this difference for a number of properties. For Roberta, only few properties yield the expected probability difference.

Pre-trained language models have not been trained on a specific semantic task. Rather, they have simply learned to predict masked tokens (or next sentences). The success of contextualized language models is rooted in their high performance when fine-tuned on a task-specific training set for a specific task. It can be argued that the fine-tuning process 'foregrounds' different aspects of linguistic knowledge that are relevant for a particular task. Thus, another way of assessing the conceptual knowledge captured by contextualized language models is to test whether the models can learn to reason over semantic phenomena when fine-tuned on a task that requires knowledge of and reasoning over semantic properties.

The Winograd Schema Challenge (WSC) (Levesque et al., 2012) has been designed to assess common sense knowledge and reasoning abilities. The task poses pronoun-resolution problems that can, at least in theory, only be solved by means of reasoning over aspects of common sense knowledge. Consider Example 21:

(21) The *trophy* doesn't fit into the brown *suitcase* because it is too **large**.

The task is to choose the correct co-referent of the pronoun *it* from the two candidate noun phrases *trophy* and *suitcase*. If a model knows that suitcases tend to be used as containers, it should be able to leverage this knowledge and correctly assign *it* to *trophy*. The task has been re-designed to be suitable for a multiple-choice set-up commonly used for language models by replacing the pronoun with a gap. The language model can then be fine-tuned on a binary classification task in which it learns to distinguish correct from incorrect solutions for filling the gap (Example 22):

(22) The *trophy* doesn't fit into the brown *suitcase* because the __ is too **large**.

a. The *trophy* doesn't fit into the brown *suitcase* because the trophy is too **large**. (correct)

b. The *trophy* doesn't fit into the brown *suitcase* because the suitcase is too **large**. (incorrect)

The fine-tuning process should foreground the relevant information for reasoning over common sense problems if it is indeed captured by the language model. Existing Winograd datasets, however, do not allow for a targeted analysis of different aspects of semantic knowledge. Therefore, I present a study that uses the diagnostic dataset in a Winograd-style challenge. This is done by means of a template-based approach in which positive and negative examples of a property are used as co-reference candidates (Example 23):

(23) John prefers the *seagull* over the *penguin* because the __ can **fly**.

The template-based approach means that the sentences seen by the model are possibly unconventional and may sound unnatural. However, a language model that has sufficient knowledge should not be distracted by this. The artificial nature of the task has the advantage that we have control over what is expressed in the sentences.

Having control over the Winograd sentences is particularly relevant, since recent research has shown that high performance on the Winograd Schema Challenge (WSC) as well as other semantic tasks may be the result of spurious correlations in the data, rather than the models' ability to reason over the target information (e.g. Elazar et al., 2021; Abdou et al., 2020; Poliak et al., 2018). As a response to this criticism, Sakaguchi et al. (2020) have proposed a WSC-style dataset (called Winogrande) in which they attempted to remove such spurious correlations by identifying and discarding instances with obvious lexical associations. State of the art models performed less highly on this dataset, indicating that (1) the dataset may indeed be less biased and (2) that the common-sense reasoning abilities of state of the art models are not as impressive as they seemed. To explore what type of knowledge language models trained on the Winogrande training set capture, I present an evaluation of the fine-tuned models using a Winograd-style challenge set constructed around the properties and concepts in the diagnostic dataset. This evaluation set has been created by means of a template-based approach. One advantage of this template-based approach is that it allows for systematic exploration of what the models trained on the Winogrande set capture (Section 10.4).

The results of the Winograd-style property challenge indicates that the fine-trained models under investigation perform barely above a random, chance-based baseline across most properties. To determine whether the models rely on superficial features rather than property knowledge, systematic variation in the templates used for generating the challenge were exploited. The results of this analysis show indications that the models fall back on using marked and unmarked discourse structures for their decisions rather than reasoning over properties. These results in combination with the initial insights based on the cloze tasks lead to the following conclusions: If property knowledge in encoded in the models, its signal is most likely too weak to override unconventional discourse structure in the sentences.

This chapter is structured as follows: I first give present the diagnostic dataset and explain how it can be used to challenge contextualized models (Section 10.2). Section 10.3 presents the results of the cloze task for pre-trained contextualized models and Section 10.4 the results of the Winograd-style challenge for fine-tuned models. The results of both studies are discussed in Section 10.5.

10.2 Diagnostic data

Both tasks use semantic property dataset introduced in Part III of this thesis. Originally, it was designed to 'diagnose' semantic property knowledge in context free embeddings by means of a probing task in which a classifier has to distinguish positive from negative examples of a property. The positive and negative examples have high semantic similarities due to shared taxonomic categories and can thus be expected to have similar embedding representations. They are, however, still distinguishable by the semantic property in question (e.g. *seagull*

Category	Property	n pos	n neg
complex	dangerous	77	60
	used_in_cooking	100	45
function/ action	fly	65	104
	roll	55	42
	swim	101	47
part	wheels	78	27
	wings	82	84
taxonomic	lay_eggs	75	70
perceptual (temperature)	cold	70	24
	hot	103	43
	warm	133	36
perceptual (color)	black	90	53
	blue	60	110
	green	94	69
	red	92	69
	yellow	43	88
perceptual (material)	made_of_wood	100	45
perceptual (color)	round	103	20
	square	90	22
perceptual (taste)	juicy	92	64
	sweet	99	64
gender (control)	female	152	208

Table 10.1: Overview of the property types and label distribution in the diagnostic dataset.

and *penguin* can be distinguished by **fly**). Thus, the dataset allows for a systematic analysis of semantic properties and avoids obvious lexical associations. An overview of properties and the number of positive and negative examples per property is shown in Table 10.1. As in the diagnostic experiments presented in Chapter 8, the property **female** is used as a control condition. Information about gender can be expected to be encoded systematically. If the models cannot perform well on the control property, this may be an indication that the interpretability method used is not able to detect semantic property information.

10.3 Study 1: Two Cloze-Task Challenges

The first study presented in this chapter examines the information captured by bi-directional contextualized language models on the basis of pre-training. Pre-training refers to the process of creating a language model by means of masked token prediction and, in the case of Bert, next sentence prediction (for a more detailed explanation, refer to the core concepts explained in Section 1.2.2 of Chapter 1). It is important to note that pre-trained language models have only been trained on predicting tokens (and sentences) given a particular context. Their pre-training does not include specific tasks that highlight semantic property information.

Nevertheless, if contexts do indeed reflect property-information, a simple cloze-task should at least show initial indications that this is the case.

The experiments presented in this section rely on comparing token probabilities given a particular sentence (e.g. The *lemon* is [mask]. v.s. The *sea* is [mask].). It should be noted that such experiments require a number of methodological choices that have implications for the conclusions drawn from them. The experiments presented here constitute an initial approach and should be seen as a first step, rather than an exhaustive analysis.

10.3.1 Method and data

In this section, I introduce two variants of a cloze-task: concept prediction given the property (Example 24a) and property prediction given the concept (Example 24b).

(24) a. The [mask] is **yellow**.
b. The *lemon* is [mask].

Firstly, I present the template sentences used for both variants. Secondly, I outline the extraction of token probabilities. Thirdly, I introduce a random baseline against which probability differences can be interpreted.

Templates

Concept prediction To compare probabilities predicted by the language models, I embed the positive and negative examples in the diagnostic dataset in sentences that evoke the property using templates. The sentences in Example 25 illustrate this idea:

(25) a. The [mask] flew.
b. The seagull flew.
c. The penguin flew.

To provide conditions that are equivalent to the masked language modeling scenario, I use the sentence-separation tokens used when pre-training the language models. Thus, for Bert, the template then becomes: [CLS] The [mask] can fly [SEP].

When creating templates for both variants of the challenge, the following aspects should be considered: Properties differ with respect to how they relate to concepts (e.g. color attributes v.s. parts). To create templates that are close to natural language, I embed properties and concepts in slightly different templates by choosing different predicates for different property-types, as shown below:

perceptual : [CLS] The [MASK] is yellow. [SEP]

perceptual (material) : [CLS] The [MASK] is made of wood. [SEP]

complex : [CLS] The [MASK] is dangerous. [SEP]

complex (used_in_cooking) : [CLS] I used the [MASK] to cook something. [SEP]

part : [CLS] The [MASK] has wings. [SEP]

taxonomic : [CLS] The [MASK] lays eggs. [SEP]

action : [CLS] The [MASK] flew. [SEP]

gender (control) : [CLS] The [MASK] showed herself. [SEP]

The templates shown above target mentions of concepts that refer to specific instances (e.g. *the lemon*) rather than general statements (e.g. *all lemons*). This approach is in line with the hypotheses about property-mentions in corpora presented in Chapter 3. In general, template-based approaches offer a variety of choices. For instance, it could be considered to elicit subtle information about quantifier relations (ALL v.s. SOME) between properties and concepts. It is, however, highly questionable whether such templates could elicit probability differences, as it has been shown that cloze templates involving negation hardly lead to correct predictions (Ettinger, 2020). In this exploratory approach, I use the simple templates presented above to get first insights. If the language models do not show clear probability differences based on the templates in this approach, it is unlikely that more subtle distinctions can be elicited.

Property-prediction The concept-evoking sentences presented above provide semantically bleached contexts that do not contain a high degree of information. The semantic properties in the diagnostic dataset can apply to a wide variety of concepts and thus do not provide much information to the language model. It can be expected that the probabilities for the masked slot will thus remain low and might be influenced by noise. To make the results more robust, I add a variation of the task in which the concept is provided, but the property is masked as shown in Example 26. If the model captures information about the property **fly** for the word *seagull*, but not for the word *penguin*, this should be reflected in the probabilities it assigns to *flew*.

(26) a. The [concept] [mask].
b. The seagull [mask].
c. The penguin [mask].

The two variants of the cloze task assess slightly different types of association: The concept prediction variant (i.e. the property is given) assesses how strongly a property is associated with a concept. The property prediction variant assesses how strongly specific concept evokes the target property. Both scenarios should yield higher probabilities for positive examples of a property than for negative examples or a property.

Probability Extraction and Normalization

For each property-evoking sentence, I extract the probability distribution over the model vocabulary (i.e. the softmax probabilities of the final masked token prediction layer) for the masked token. I extract the probability assigned to the respective target token (i.e. the word

expressing the concept or property). If the model captures property-evidence, the positive examples should yield higher probabilities than the negative examples.

When extracting the probability p of a token for a masked position in a sentence, it should be kept in mind that the probability of the token depends on the context, but also on the probability of the token independently of the context as pointed out by Kurita et al. (2019). For instance, the words *seagull* and *penguin* may have different probabilities independent of their specific contexts (due to their frequency and distribution in the training data). This factor could interfere with the probability comparison.

To account for this type of 'prior' probability (p_{prior}) of a word independent of its context, I use a normalization strategy suggested by Kurita et al. (2019): I use a minimal context to approximate the prior probability of the target word (either an example concept or the word expressing the property). As shown in Example 27, I mask the property- or concept-specific part of the context and extract the probability of the target word given the minimal context.

(27) a. The [mask] flew. (original)

b. The [mask] [mask]. (minimal context)

The second sentence in the example masks the property in addition to the concept and thus poses a minimal context. The probability of the masked concept in the second sentence thus approximates the tendency of the language model to predict the concept independent of the property-evoking context. I compute the normalized probability as $log(\frac{p}{p_{prior}})$ following Kurita et al. (2019).

Interpretation of Probability Differences

In essence, a language model that captures property-information should assign higher probabilities to the positive examples of a property than to negative examples of the property (concept prediction variant). Likewise, it should assign higher probabilities to the property when given a sentence containing a positive example of the property than when given a negative example (property prediction variant). The positive and negative examples in the diagnostic dataset allow for such a probability comparison. I establish the probability difference for a specific property by calculating the difference between the mean probability of all positive examples and the mean probability of all negative examples.

The core difficulty of this task setup is to establish whether an observed difference between positive and negative examples is likely to be meaningful, or whether it could be due to chance. In order to establish this, I apply the following, high bar: I assign random labels (i.e. positive or negative) to all concepts in a property dataset. For instance, the positive and negative examples of the property **fly** are randomly assigned to either the positive or negative class (e.g. *airplane* and *penguin* may receive a negative label, *bus* and *bee* may receive a positive label). The resulting randomized dataset should not lead to a meaningful difference between the mean probability of all randomly assigned positive examples and the mean probability of all randomly assigned negative examples. If the difference observed on the real label distribution is meaningful, it should be higher than the difference in the random condition. I repeat the random label assignment 100 times and use the highest difference as the baseline.

This high bar entails the following risks: One of the 100 random label assignments could happen to follow the distribution of the original dataset (a highly unlikely situation). To ensure that this is not the case, all random label assignments have been checked. A second risk is that the dataset contains noise and the random label assignment happened to distribute noisy examples in such a way that it leads to a justified, higher probability difference. While this scenario is possible, it can still be expected that a clear difference between the majority of examples should be robust enough result in a high score given a few noisy examples.

10.3.2 Experimental Setup and Results

In this Section, I present the results of the masked token prediction task. The section is structured as follows: Firstly, I describe the pre-trained contextualized models used to perform the task. Secondly, I present a validation of the templates. The purpose of the validation is to show that two cloze task variants (concept prediction and property prediction) do indeed lead to plausible sentence completion. Thirdly, I present the results of the two task variants and conduct a comparison.

Contextualized Models

I experiment with two types of transformer-based bidirectional language models: Bert (Devlin et al., 2019) and Roberta (Liu et al., 2019). While the models share the same architecture, they differ with respect to pre-training: Bert is trained on masked token prediction and next sentence prediction, while Roberta is only trained on the former. Both pre-trained models come in two sizes (Bert-base-uncased and Bert-large-uncased; Roberta-base and Roberta-large).[1] I use both variants of both models in the cloze task experiments.

Template Validation

To illustrate that the template sentences are in principle suitable for this analysis, I have retrieved the top five predictions for the masked slots for both tasks. The tokens with the top probabilities should be appropriate concepts or properties. This inspection can indicate whether the templates are suitable for the task.

Table 10.2 shows the top five tokens predicted for the masked slot in the concept prediction variant for Bert-large and Roberta-large. It can be observed that for all properties, there are plausible candidates among the predicted concepts. However, the candidates also contain noise that is, in many cases, quite obviously the result of superficial lexical associations rather than a deep, semantic understanding of the sentence (e.g. **wings**: *male*, *female*, *species* in Bert-large). Roberta-large seems to produce comparatively more abstract and noisy predictions (e.g. **sweet**: *life*, *venge*, *emption*, *love*, *it*) than Bert-large (e.g. **sweet**: *blood*, *water*, *coffee*, *food*, *smell*). The mean prediction probabilities of the top five tokens are below 0.14 (Bert-large) and 0.02 (Roberta-large). To validate the property-prediction variant of the task, I also extract the top-five predictions for the properties given the templates containing the concepts.

[1] All four models were accessed via the huggingface transformers library `https://huggingface.co/transformers/`

	bert-large-uncased		roberta-large	
	top5	prob.	top5	prob.
square	base body aperture shape tower	0.05	circle square world Square box	0.04
warm	night air day sun room	0.14	weather sun water it air	0.07
black	head underside abdomen iris apex	0.1	author character color writer background	0.03
red	iris mouth color aperture underside	0.05	color background code font logo	0.02
dangerous	man world situation future woman	0.04	this world This it news	0.02
wings	male female species animal body	0.11	bird sun cat future devil	0.02
sweet	blood water coffee food smell	0.05	life venge emption love it	0.03
hot	water sun air day room	0.12	water it sun weather world	0.02
juicy	food fruit meat stuff pie	0.08	story this one This stuff	0.05
green	color colour iris underside bark	0.07	grass color world sky tree	0.03
made_of_wood	building roof frame structure statue	0.05	house furniture chair car table	0.02
blue	sky background ground iris band	0.07	color background sky code text	0.04
yellow	abdomen underside iris head body	0.1	color background text code image	0.02
cold	air room night water wind	0.13	world it weather water winter	0.04
round	aperture nest shell body fruit	0.16	world Earth universe circle earth	0.07
wheels	car boat ship vehicle carriage	0.05	bus train car future world	0.04
lay_eggs	female male hen species pair	0.2	chicken hen she also cow	0.03
roll	credits cameras camera thunder wheels	0.09	dice eyes heads we I	0.08
swim	water world man fish room	0.04	he she I they we	0.07
fly	bullets sparks ball fire knife	0.05	they he it she I	0.07
used_in_cooking	time microwave money knife heat	0.05	microwave oven stove time pot	0.08
female	woman girl queen witch lady	0.13	she then finally never woman	0.08

Table 10.2: Top five predicted tokens and their mean probabilities (raw) for each property-evoking sentence in Bert large and Roberta large.

Table 10.3 shows the resulting predictions for Bert-large and Roberta-large. In many cases, the target-property is among the top five predicted tokens. For properties where this is not the case, the predictions are often related to the target property or can at least be explained (e.g. *crashed* for **fly**, *landed* for **wings**). It is important to note that the task does not require

	bert-large-uncased		roberta-large	
	top5	prob.	top5	prob.
square	empty open gone blank ringing	0.05	empty full available closed broken	0.03
warm	black white red gone yellow	0.05	optional black on white required	0.02
black	black nocturnal edible dead gone	0.03	dead gone back here over	0.02
red	edible white black yellow gone	0.04	gone dead delicious red optional	0.01
fly	##s vol press family ##ch	0.03	s </s> here crashed ling	0.01
dangerous	dead gone empty over loaded	0.02	dead back here gone loaded	0.03
wings	wings disappeared arrived gone died	0.04	arrived landed returned died crashed	0.06
sweet	edible black white yellow green	0.04	delicious gone here red ripe	0.02
hot	empty white good black delicious	0.03	on empty delicious optional broken	0.02
used_in_cooking	do make cook eat cut	0.08	make cook build do cut	0.08
juicy	edible black white yellow green	0.04	delicious here gone red ripe	0.02
green	edible black white green yellow	0.04	dead delicious optional gone here	0.01
made_of_wood	wood steel oak iron bamboo	0.08	wood steel metal aluminum plastic	0.08
blue	edible black gone white empty	0.04	gone dead back broken pending	0.02
yellow	white edible yellow black good	0.04	dead gone delicious here good	0.02
roll	stopped ##s coaster flickered shook	0.02	s coaster ball list </s>	0.02
female	up him interest me her	0.12	up me her off him	0.13
cold	empty black cold good white	0.03	coming empty gone broken here	0.02
round	empty white black edible gone	0.04	delicious optional empty gone here	0.02
wheels	stopped disappeared arrived wheels no	0.05	arrived stopped wheels crashed changed	0.05
lay_eggs	eggs down low still dormant	0.20	eggs down low dormant egg	0.15
swim	##s vol press said bass	0.02	bass ex tuna s </s>	0.01

Table 10.3: Top five predicted tokens and their mean probabilities (raw) for each concept-evoking sentence in Bert large and Roberta large.

the correct answer to be among the top-n predictions. Rather, what is important is that the probabilities reflect a difference between positive and negative examples of a concept. Overall, it can be concluded that the templates are suitable for both variants of the cloze task.

	bert-base-uncased	bert-large-uncased	roberta-base	roberta-large
wheels	**0.8599**	-0.2450	-0.2045	-0.9568
used_in_cooking	**0.7189**	**0.6289**	-0.0631	-0.0898
blue	**0.3682**	**0.5814**	-0.2181	**0.0430**
wings	**0.2060**	**0.0680**	-0.2580	**0.0610**
made_of_wood	**0.1174**	-0.2185	-0.1804	-0.2069
green	-0.1442	**0.2418**	-0.3376	-0.4528
red	-0.3955	**0.2297**	-0.0106	-0.2092
roll	-0.1586	-0.0248	-0.4178	**0.0682**
black	-0.0214	-0.3922	-0.3899	-0.4054
lay_eggs	-0.0638	-0.0782	-0.3494	-0.2541
round	-0.0967	-0.4509	-0.2027	-0.5405
dangerous	-0.1107	0.1426	-0.0720	-0.2032
warm	-0.2515	-0.3247	-0.4561	-0.9085
swim	-0.3011	-0.3299	-0.4365	-1.4518
yellow	-0.4086	-0.2713	-0.1660	-0.8454
sweet	-0.4299	-0.9187	-0.4289	-0.5302
cold	-0.5777	-0.8095	-1.2415	-1.1395
hot	-0.6640	-1.3849	-0.5225	-0.7626
fly	-0.6919	-0.5358	-0.0727	-1.3138
juicy	-0.8733	-1.3109	-0.5426	-0.7680
square	-1.6022	-1.4540	-1.2632	-1.5935
female	**0.4874**	**0.5506**	-0.1687	-0.4488

Table 10.4: Summarized results of the concept prediction task for both variants of the Bert and Roberta model. The scores indicate the difference between the observed probability differences in the true label distribution and the maximum difference out of 100 randomized label distributions. A difference above 0 indicates that the models show clear differences between positive and negative examples.

Concept and Property Prediction

In this section, I present the results of the concept and property prediction tasks. To establish whether the difference is likely to be meaningful, I compare the observed difference on the task to the maximal difference observed when using random label assignment.

Concept prediction Table 10.4 shows the summarized results of the masked concept prediction variant. The positive values indicate that the difference between the mean probability of positive examples and mean probability of negative examples is higher then the difference observed on the random baseline (values shown in bold). Seven properties outperform the random baseline in at least one of the four models. Out of the seven properties, three outperform the random baseline in one of the Roberta models. Roberta-base does not outperform the random baseline for any of the properties. Both Bert models outperform the random baseline for the control property **female** by a clear margin. Neither of the Roberta models outperform the random baseline for **female**.

	bert-base-uncased	bert-large-uncased	roberta-base	roberta-large
used_in_cooking	**2.3578**	**2.4782**	**2.7690**	**1.8501**
wings	**1.6188**	**1.3682**	**0.6248**	-0.0578
hot	**0.7840**	**0.8275**	**0.0271**	**0.1887**
wheels	**0.7061**	**1.6392**	-0.1039	-0.6507
green	**0.6025**	**0.7346**	**1.2773**	**1.2120**
lay_eggs	**0.3598**	**0.3520**	-1.4953	**0.1576**
warm	**0.3383**	**0.1806**	**0.1170**	**0.2956**
dangerous	**0.2852**	**0.1376**	-0.6905	-0.7087
fly	**0.1102**	**0.5100**	-0.8065	-0.7780
made_of_wood	**0.1076**	**0.3995**	**0.1014**	**0.5360**
blue	**0.0691**	-0.0576	-0.1578	**0.1311**
yellow	**0.0069**	-0.1509	-0.7169	-0.6117
sweet	-0.0147	**0.3960**	**0.8809**	**0.6800**
juicy	-0.2231	**0.1846**	-0.3222	-0.8769
swim	-0.6935	**0.5897**	-0.5413	-0.3684
cold	-0.0283	-0.7279	**0.6181**	-0.2748
red	-0.2063	-0.1814	-0.4320	-0.2236
round	-0.4573	-0.2933	-1.0528	-0.3465
black	-0.5344	-0.4560	-0.7690	-0.2760
roll	-0.6580	-0.6652	-0.4178	-0.9240
square	-1.8323	-1.7282	-3.1179	-1.8037
female	**2.1705**	**1.7841**	-0.4563	-1.2669

Table 10.5: Summarized results of the property prediction task for both variants of the Bert and Roberta model. The scores indicate the difference between the observed probability differences in the true label distribution and the maximum difference out of 100 randomized label distributions. A difference above 0 indicates that the models show clear differences between positive and negative examples.

Property prediction Table 10.5 shows the summarized results of the property prediction variant. It can be seen at first glance that many more properties outperform the random baseline on this task in the Bert models (16 compared to 7). Most properties that outperform the baseline also do so by a considerably higher margin than on the concept-prediction task (2.3578 vs 0.7189 for **used_in_cooking** Bert-base). As in the concept prediction challenge, the Roberta models outperform the random baseline for fewer properties than the Bert models (8 for Roberta-base v.s. 12 for Bert-base and 8 for Roberta-large v.s. 13 for Bert-large). Both Bert models outperform the random baseline for the control property **female** with a considerable margin, but neither of the Roberta models do.

Task Comparison

The results of the two cloze tasks indicate that the prediction of a property given a context containing a concept is considerably easier for the contextualized models than the prediction of a concept given a property. A possible explanation for this behavior could lie in the difference of categorical specificity expressed by concepts and properties. Bolognesi et al.

(2020) discuss concreteness and categorical specificity (as defined by Borghi and Binkofski (2014)) as two distinct phenomena involved in abstraction. Two concrete concepts can have different levels of categorical specificity: *rocking chair* (high specificity) versus *furniture* (low specificity). Concepts with high specificity carry a high number of semantic properties and can thus apply to a comparatively small set of referents in the world. It could be argued that concepts with high categorical specificity appear in a comparatively small selection of different linguistic contexts and provide better, more informative indications for a language model than a categorically less specific concept. The properties as well as the concepts in the diagnostic datasets could be seen as concepts of different levels of categorical specificity: For instance, the property **wings** also expresses the concept *wings*. When compared to some of the concepts in the dataset (e.g. *dragonfly*, *sparrow*), it can be argued that the property has lower categorical specificity. This intuition can also be explained as follows: It is much more likely to think of *wings* when presented with the word *dragonfly* than the think of *dragonfly* when presented with the word *wings*. In this section, I explore these intuitions on the basis of a small set of model predictions in the two task variants.

	label	concept		property	
		prob_norm	prob	prob_norm	prob
wasp	pos	3.0825	0.0003	7.333	0.2898
bat	pos	2.6084	0.001	7.3964	0.3088
eagle	pos	2.0552	0.0004	8.2918	0.756
crow	pos	2.0223	0.0005	7.0214	0.2122
beetle	pos	2.0147	0.0011	6.6249	0.1428
robin	pos	1.929	0.0001	3.2144	0.0047
worker	neg	1.9073	0.0002	1.3656	0.0007
queen	neg	1.8885	0.0021	-0.1162	0.0002
moth	pos	1.8568	0.0027	6.5584	0.1336
insect	pos	1.529	0.0016	7.4471	0.3248
bird	pos	1.3928	0.005	7.2665	0.2712
kite	neg	1.3134	0.0001	6.5757	0.1359
drone	pos	1.1308	0.0002	1.0869	0.0006
clarence	neg	1.0887	0	2.6964	0.0028
monarch	neg	1.0539	0.0002	0.1709	0.0002
hen	pos	0.8963	0	5.4375	0.0435
parachute	neg	0.8913	0	-0.4467	0.0001
flea	neg	0.8031	0	4.3416	0.0146
plane	pos	0.7372	0.0008	5.8557	0.0661
nightingale	pos	0.6944	0	6.3744	0.1111

Table 10.6: Top 20 concepts with the highest normalized and raw probabilities in the concept prediction task ('concept') compared to the property prediction task ('property') for the property **wings** in Bert-base.

To explore the differences between the two task variants, I inspect the top 20 predictions for the property **wings** in the Bert-base model. Table 10.6 shows the 20 highest probabilities assigned to concepts given the property **wings**. Table 10.7 shows the 20 highest probabilities assigned to the property **wings** given the concepts in the dataset. Both tables also show

	label	property		concept	
		prob_norm	prob	prob_norm	prob
eagle	pos	8.2918	0.756	2.0552	0.0004
dragonfly	pos	7.5428	0.3574	-1.1437	0
insect	pos	7.4471	0.3248	1.529	0.0016
sparrow	pos	7.409	0.3127	-0.417	0
bat	pos	7.3964	0.3088	2.6084	0.001
gnat	pos	7.3853	0.3054	-1.1437	0
wasp	pos	7.333	0.2898	3.0825	0.0003
ant	neg	7.2672	0.2714	-1.2166	0
bird	pos	7.2665	0.2712	1.3928	0.005
damselfly	pos	7.0873	0.2267	-1.1437	0
crow	pos	7.0214	0.2122	2.0223	0.0005
owl	pos	6.8499	0.1788	0.1195	0
cabriolet	neg	6.6319	0.1437	-1.1437	0
beetle	pos	6.6249	0.1428	2.0147	0.0011
kite	neg	6.5757	0.1359	1.3134	0.0001
grasshopper	pos	6.5708	0.1352	-1.1437	0
moth	pos	6.5584	0.1336	1.8568	0.0027
falcon	pos	6.432	0.1177	-1.2053	0
nightingale	pos	6.3744	0.1111	0.6944	0
duck	pos	6.358	0.1093	-1.5784	0

Table 10.7: Top 20 concepts with the highest normalized and raw probabilities in the property prediction task ('property') compared to the concept prediction task ('concept') for the property **wings** in Bert-base.

the probabilities assigned to the property or concept in the respective other variant of the challenge. Overall, it can be observed that the probabilities assigned to concepts given the property are generally lower than the probabilities assigned to the property given the concept. The same tendency can also be observed on the level of specific concepts: For example, the probability assigned to *eagle* given the property **wings** is much lower (2.0552) than the probability assigned to **wings** given *eagle* (8.2918). This difference could be seen as a first indication that the concepts provide overall more categorically specific and thus more informative concepts than the properties.

When considering individual concepts given the property (Table 10.6), it is striking that several negative examples receive comparatively high probabilities: *queen*, *worker*, *monarch*, *kite*. All words except for *kite* trigger comparatively low probabilities in the property-prediction task. Three out of the four examples are ambiguous and have senses for which the property **wings** applies (*queen* in the sense of *bee*, *monarch* in the sense of *butterfly*, *worker* in the sense of insect). *Kite* could be seen as a vague concept with respect to the having wings: Kites do not have typical wings, but technical descriptions of kites use the term *wings*.[2] The (perhaps somewhat unpopular) senses of the three ambiguous words are not reflected by the crowd annotations. The probabilities assigned in the property-prediction task

[2]The term *wings* is mentioned in the Wikipedia article about kites `https://en.wikipedia.org/wiki/Kite` (last accessed 2021/10/27).

follow the crowd-intuitions and the arguably more popular senses (e.g. *queen* in the sense of female monarch), while the probabilities assigned in the concept-prediction reflect the senses of the words to which the property **wings** applies (e.g. *queen* in the sense of bee). It is likely that the context containing the word *wings* simply triggers representations of the less popular uses of the ambiguous words. This behavior illustrates that Bert-base does indeed use a contextualized approach; it seems to access the representation of the word *queen* that best fits the context of the word *wings*. The different representations of polysemous words may indeed approximate different senses.

When considering the probabilities predicted for the property **wings** given a sentence containing the concepts presented in Table 10.7, the following observations can be made: For several of the top-ranked concepts, there is a strikingly high difference in probabilities between the two challenges (e.g. *dragonfly*, *sparrow*, *gnat*, *damselfly*). For these concepts, the probabilities predicted given the property are strikingly low. The concepts do not rank among the top-predicted concepts in the concept prediction task. It could be argued that these concepts are categorically highly specific and provide a stronger signal to the language model than the property **wings**. In other words, the concept *dragonfly* is more likely to evoke the property **wings** while the property **wings** is unlikely to trigger the categorically highly specific concept *dragonfly*.

The property prediction challenge also assigned high probabilities to negative examples (*ant*, *cabriolet*, *kite*). All of the three examples can be explained: Parts of kites can be called wings and certain types of ants have wings. Thus, the annotations for *kite* and *ant* can be seen as inaccurate. In the context of *cabriolet*, the word *wings* can refer to a car-part (protective mud wings surrounding the wheel). It is possible that the word *cabriolet* triggered this use of the word *wings*. This observation illustrates the ability of a contextualized language model to select word representations based on context (and thus possibly approximate word senses). However, in the current task set-up, it has the disadvantage that this ability of the model complicates the interpretation of its performance on the task.

To summarize, the comparison of the top-20 probabilities predicted in the two masked token prediction tasks yielded the following insights: Firstly, it seems that the property-prediction variant of the task tends to trigger higher probabilities and could be seen as 'easier' for the language model. A likely reason for this tendency is that the concepts tend to be categorically more specific than the property. The word expressing the property is thus a salient context of the concept, but the concept is not salient for the property. This is particularly apparent for highly specific concepts (e.g. *damselfly*). Secondly, the analysis showed that Bert does indeed capture context-specific representations of words.

10.4 Study 2: Winograd-Style Challenge

The second study presented in this chapter constitutes an exploration of the reasoning abilities of contextualized models fine-tuned on the Winogrande training set (Sakaguchi et al., 2020).[3] The Winograd Schema Challenge (Levesque et al., 2012) was designed to assess the

[3]The study was conducted in collaboration with Sanne Hoeken and Piek Vossen. Sanne Hoeken generated the Winograd-style dataset and implemented the experimental set-up. The section is based on a first draft of a paper

common sense reasoning abilities of NLP systems by means of a pronoun resolution task (see Example 28):

(28) The *trophy* doesn't fit into the brown *suitcase* because it is too **large**.

The original dataset was criticized for containing biases as models started to reach close to human performance (Sakaguchi et al., 2020). As an alternative, Sakaguchi et al. (2020) proposed Winogrande, a larger dataset that has specifically been checked for potential biases models can exploit, such as obvious lexical associations that can point towards the correct answer without requiring deeper reasoning.

The goal of the study presented in this chapter is to explore whether models fine-tuned on the Winogrande training set learned how to reason over semantic properties. The Winogrande test set does not contain systematic information about what type of common sense knowledge it captures. We study semantic property knowledge and reasoning abilities in fine-tuned models systematically by means of a Winograd-style challenge dataset constructed around the properties and concepts in the diagnostic dataset.

To gain systematic insights into the information about semantic properties captured by language models, the properties and concepts in the diagnostic dataset have to be embedded into WSC-style sentences. In an initial exploration, we approached this problem by searching for examples in the Winogrande dataset that contain the semantic properties and concepts from the diagnostic dataset. We searched for examples in the Winogrande test set that could provide indications about property knowledge. Initially, we used the following strategy: We searched for sentences that contained a property, a positive example concept from the property dataset and a negative example from the property dataset. As this search did not return examples, we loosened the search criteria as follows: We searched for words in the Winogrande dataset that have a high cosine similarity to concepts and properties from the diagnostic dataset.[4] The searches only returned a small set of noisy examples (38) that would have required manual filtering.

To exploit all properties and concepts in the diagnostic dataset, we opt for a template-based approach. We automatically generate Wingrad sentences by embedding the properties and concepts in a set of different templates. While template-based approaches can be criticized for resulting in potentially unnatural-sounding sentences, they have the advantage of allowing for systematic variations. Such variations can be used to explore the behavior of a model.

10.4.1 Method and Data

In this section, we present the template-based approach used to construct Winograd-style examples around properties and concepts from the diagnostic dataset. A consequence of the template-based approach is that all test-instances have the same syntactic structure. We present how we can exploit systematic variations in the templates to explore whether the fine-tuned models are likely to base their decision on actual property knowledge or whether

written in collaboration with Sanne Hoeken and Piek Vossen.

[4]The cosine similarity was measured on the basis of the GoogleNews embedding models `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?resourcekey=0-wjGZdNAUop6WykTtMip30g`. We experimented with different similarity thresholds.

they fall back on different structural features of the template sentences. It should be noted that our template-based dataset is only used for evaluation. We experiment with the models fine-tuned on the Winogrande dataset by Sakaguchi et al. (2020).

Winograd Templates

To test property knowledge in a contextualized task, we embed our concept pairs and properties in generic sentences based on a variety of templates. The main motivation behind picking generic sentences is that they are compatible with any combination of properties and concepts. For example, we embed the combination of *penguin*, *seagull* and **fly** in a generic sentence (Example 29). Such templates provide semantically bleached contexts and thus have the additional advantage that they are unlikely to introduce unwanted semantic associations or biases.

(29) John prefers the *seagull* over the *penguin* because the _ can **fly**.

We created a total of eight different templates to introduce a variety of different, but still generic sentences. A full overview can be seen in Table 10.8. Each combination of concept pair and property is assigned to one of the templates at random. This method eventually resulted in a dataset consisting of 106,654 instances containing 21 different semantic properties and the control property, 1,337 different positive concepts and 906 different negative concepts.[5]

Template-based approaches have the disadvantage that the sentences presented to the language model are noticeably artificial and the combination of concepts and properties might be unexpected for a language model in several instances. However, they still adhere to appropriate syntax and morphology and contain grammatical constructions the model is likely to have seen during pre-training and fine-tuning.

The templates we use have the advantage of a minimal semantic context. Thus, the chances of inferring the correct answer based on other context than just the target concepts and properties is low. Furthermore, the templates allow for testing the effect of systematic context variations, which can help us to detect whether the models do exploit spurious correlations unrelated to the target information.

Systematic Template Variations

To test whether the templates introduce spurious correlations that the models can exploit, we (1) consider factors already present in the templates and (2) introduce an additional variation with respect to marked and unmarked discourse structure.

The templates introduced in the previous section differ with respect to the following aspects: **Predicates**: Each of the 8 templates is constructed around a different predicate. **Syntactic structure**: Five of the templates place the concepts in object position and use a person (expressed by a generic name) in subject position. Two of the templates place one

[5]The full dataset can be found in the following Github repository, together with the code used for our experiments on the fine-tuned models: `https://github.com/SanneHoeken/diagnostic_dataset_experiments`

John prefers the [concept] over the [concept] because the _ [PROP]. *John prefers the seagull over the penguin because the _ can fly.*
John replaced the [concept] by the [concept] because the _ [PROP]. *John replaced the orange by the lemon because the _ is yellow.*
John chose the [concept] instead of the [concept] because the _ [PROP]. *John chose the peach instead of the wintergreen because the _ is sweet.*
John likes the [concept] but not the [concept] because the _ [PROP]. *John likes the currant but not the chipotle because the _ is sweet..*
The [concept] is better than the [concept] because the _ [PROP]. *The broccoli is better than the tyre because the _ is green.*
The [concept] is worse than the [concept] because the _ [PROP]. *The bear is worse than the parrotfish because the _ is blue.*
John has the [concept] and the [concept], the _ [PROP]. *John has the beer and the quesadilla, the _ is cold..*
There is the [concept] and the [concept], the _ [PROP]. *There is the oregano and the plane, the _ is used in cooking.*

Table 10.8: Overview of Winograd templates.

of the concept options in subject position. One template places the candidates in subject complement position. We evaluate each variation separately to determine whether the two factors have an impact on performance.

The structure of the templates allows for varying the sequence in which the candidate concepts are mentioned. For instance, the example sentence introduced in the previous section can be modified by switching the positions of *seagull* and *penguin*, as shown in Example 30. The modified version of the sentence constitutes an uncommon and unexpected structure because the concept placed in the focus position (*penguin*) is not the concept selected by the property. We consider this manner of presenting information a **marked discourse structure**. If a model has sufficient information about semantic properties, it should be able to recognize the connection between the property and the concept despite the marked structure. In contrast, if a model does not receive a sufficiently strong signal from the property-concept combination, it is likely to rely on other signals and will thus be fooled by the unconventional structure.

(30) John prefers the *penguin* over the *seagull* because the _ can **fly**.

When considering the marked and unmarked variants (shown in Table 10.9), it can be noticed that not all variations appear to be marked equally strongly. We can observe that the patterns using the predicates *prefers* and *chose* have strongly marked variants. In contrast, the variants of the patterns using *has* and *there is* are almost equally expected. We test the effect of marked and unmarked discourse structures by evaluating both possible structures for each instance in the test data.

Unmarked	Marked
John prefers the [POS] over the [NEG] because the _ [PROP]. *John prefers the seagull over the penguin because the _ can fly.*	John prefers the [NEG] over the [POS] because the _ [PROP]. *John prefers the penguin over the seagull because the _ can fly*
John replaced the [NEG] by the [POS] because the _ [PROP]. *John replaced the orange by the lemon because the _ is yellow.*	John replaced the [POS] by the [NEG] because the _ [PROP]. *John replaced the lemon by the orange because the _ is yellow.*
John chose the [POS] instead of the [NEG] because the _ [PROP]. *John chose the peach instead of the wintergreen because the _ is sweet.*	John chose the [NEG] instead of the [POS] because the _ [PROP]. *John chose the wintergreen instead of the peach because the _ is sweet.*
John likes the [POS] but not the [NEG] because the _ [PROP]. *John likes the currant but not the chipotle because the _ is sweet.*	John likes the [NEG] but not the [POS] because the _ [PROP]. *John likes the chipotle but not the currant because the _ is sweet.*
The [POS] is better than the [NEG] because the _ [PROP]. *The broccoli is better than the tyre because the _ is green.*	The [NEG] is better than the [POS] because the _ [PROP]. *The tyre is better than the broccoli because the _ is green.*
The [POS] is worse than the [NEG] because the _ [PROP]. *The parrotfish is worse than the bear because the _ is blue.*	The [NEG] is worse than the [POS] because the _ [PROP]. *The bear is worse than the parrotfish because the _ is blue.*
John has the [POS] and the [NEG], the _ [PROP]. *John has the beer and the quesadilla, the _ is cold.*	John has the [NEG] and the [POS], the _ [PROP]. *John has the quesadilla and the beer, the _ is cold..*
There is the [POS] and the [NEG], the _ [PROP]. *There is the oregano and the plane, the _ is used in cooking.*	There is the [NEG] and the [POS], the _ [PROP]. *There is the plane and the oregano, the _ is used in cooking.*

Table 10.9: Overview of marked and unmarked Winograd templates.

10.4.2 Experimental Setup and Results

This section presents the fine-tuned models under investigation and the results of evaluating the fine-tuned models trained on Winogrande on our Winograd-style task. To gain deeper insights into the model decisions, we explore the effects of systematic variations in the Winograd templates.

Contextualized Models

We evaluate existing fine-tuned models on our template-based dataset. Sakaguchi et al. (2020) have fine-tuned the large variants of both models on the training split of the Winogrande dataset. The fine-tuned models have been trained on a multiple-choice task: The model is

	ft-Winogrande		pt (baseline)	
Dataset	**Berta**	**Roberta**	**Bert**	**Roberta**
templates	0.562	0.664	0.497	0.515
Winogrande	0.649	0.791	0.568	0.559

Table 10.10: Accuracy of contextualized language (Bert-large-uncased and Roberta-large) models on the Winograd-style property challenge and the Winogrande dataset. Results are shown for the fine-tuned models (trained on the Winogrande training set) and the pre-trained models (baseline).

given two versions of a Winograd sentence as input. In each version, the blank is filled by one of the referent candidates. The [CLS] token is then used to classify the sentence candidates as correct or incorrect. If the models do indeed acquire the ability to reason over common sense knowledge in the training process, they should also succeed on our template-based Winograd-style challenge. We use the fine-tuned models provided by Sakaguchi et al. (2020)[6] and test them on our template-based dataset.

To test whether the fine-tuning process does indeed help models to identify the right information and predict the correct referent, we compare the fine-tuned models to their pre-trained variants (bert-large-uncased and roberta-large). It is possible to use pre-trained models without fine-tuning by using a masked token prediction task to fill in the gap (e.g Ettinger, 2020). In this set-up, the gap is replaced by a masked token ([mask]). We extract the probabilities of the two candidates given the sentence and chose the candidate with the higher probability. We use the pre-trained models as a baseline.

Winograd-Style Task

Table 10.10 shows the overall accuracy of Bert and Roberta on our template-based dataset. Both fine-tuned models perform above chance level and considerably lower than on the Winogrande dataset (performance taken from the Winogrande leaderboard[7]). Both fine-tuned models perform higher than their pre-trained variants on the template-based dataset and on the original Winogrande dataset.[8] As on the original dataset, fine-tuned Roberta outperforms fine-tuned Bert. In the remainder of this section, we focus on the analysis of the fine-tuned models only.

Since our property-concept dataset allows for a systematic exploration of different semantic properties, we present the accuracy per property and property category in Table 10.11. We see that both models perform in similar ranges for most properties and property-categories. The only striking effect we can observe is that Roberta performs almost perfectly (accuracy of 0.946) for the control property **female**. For four properties, Roberta performed above 0.70: **hot**, **sweet**, **cold**, **blue**. Bert cannot go beyond an accuracy score of 0.652 (for the property **hot**) and performs comparatively low on the control property **female** (0.563).

[6]We have downloaded the fine-tuned models from `https://winogrande.allenai.org/`

[7]`https://leaderboard.allenai.org/winogrande/submissions/public`

[8]The scores of the pre-trained models for Winogrande are taken from Klein and Nabi (2021).

Category	Property	Bert	Roberta	Supp.
complex	dangerous	0.579	0.606	4620
	used_in _cooking	0.590	0.631	6890
function/action	fly	0.558	0.608	6760
	roll	0.528	0.586	2310
	swim	0.521	0.512	4747
perceptual (taste)	juicy	0.522	0.586	5888
	sweet	0.579	0.725	6336
part	wheels	0.564	0.697	2106
	wings	0.576	0.677	6888
perceptual (temperature)	cold	0.613	0.744	1680
	hot	0.652	0.771	4429
	warm	0.516	0.615	4788
taxonomic	lay_eggs	0.549	0.524	5250
perceptual (color)	black	0.553	0.675	4770
	blue	0.558	0.708	6600
	green	0.557	0.654	6486
	red	0.554	0.649	6348
	round	0.562	0.562	2060
	square	0.549	0.657	1980
	yellow	0.601	0.641	3784
perceptual (material)	made_of_wood	0.534	0.691	4500
gender (control)	female	0.563	0.946	7434

Table 10.11: Accuracy scores of the fine-tuned models on the Winograd-style property challenge per semantic property.

If a fine-tuned model learned to reason over semantic properties, we would expect that the pre-trained variant of the model also captures information about the property in question (at least to some degree). The results of the cloze-tasks presented Section 10.3 should be in line with the results of the Winograd-stype property challenge. For three out of the four highly performing properties, the pre-trained variant of the Roberta model has performed above a random baseline in at least one of the cloze tasks presented in Section 10.3 (**hot**, **sweet**, **blue**). It is striking that the pre-trained Roberta model did not perform above the random baseline for the control property **female** in any of the cloze variants presented in Section 10.3. This discrepancy between pre-trained and fine-tuned models poses the question of whether the fine-tuned Roberta model could have learned to reason over gender information. An alternative explanation for the exceptionally high performance of the fine-tuned Roberta model is that the model learned to exploit biases in the training data.

Effects of Template-Variations

In this section, we test whether systematic variations in the templates have an effect on the results. We explore the impact of the predicates, the syntactic structures and different

discourse structures. It should be noted that templates were assigned to triples of a property, a positive example, and a negative example at random. This has the consequence that the comparisons between different templates do not necessarily contain exactly the same property-concept combinations. In future work, it could be considered to compare the behavior of the models given minimal pairs (i.e. equivalent property-concept combinations that only differ in terms of a specific template variation).

Table 10.12 shows the performance of the two models with respect to the **predicates** around which the templates have been constructed. While most templates show similar performance, we can observe that *has _ and _* and *is _ and _* seem to lead to slightly higher performance than the other predicates.

predicate	Bert	RoBerta	support
_ is better than _	0,508	0,596	13022
_ is worse than _	0,514	0,592	13274
chose _ instead of _	0,523	0,648	13433
has _ and _	0,686	0,771	13298
there is _ and _	0,689	0,775	13416
likes _, but not _	0,520	0,653	13438
prefers _ over _	0,524	0,649	13296
replaced _ by _	0,528	0,628	13477

Table 10.12: Mean accuracy score of contextualized language models on the template dataset with respect to the different predicates.

Table 10.13 shows the results of the two models with respect to the **syntactic position** of the candidate concepts in the templates. The models perform slightly higher for templates with the concepts in object position than in subject position. The highest performance is reached by concepts in subject-complement position. Subject complement position is only the case for one type of template using the predicate *is _ and _*. The difference may also be caused by word order rather than syntax: In terms of word order, the subject-complement template is similar to the object-position template.

Syntactic position	**Bert-ft**	**RoBerta-ft**	**Support**
object	0.556	0.670	66942
subject	0.511	0.594	26296
subject complement	0.689	0.775	13416

Table 10.13: Mean accuracy score of contextualized language models on the templates with respect to syntactic structure.

When considering the impact of **discourse structure** we found that overall, unmarked discourse structure leads to considerably higher performance than marked structure (0.76 vs 0.37 for Bert and 0.80 vs 0.53 for Roberta). For marked discourse structure, the models perform clearly below chance level. When considering the variations by template type (Table 10.14), we can see that model performance follows the expected pattern for all cases except *_is worse than _*. The two predicates *_has and _* and *is _ and _* show a much smaller difference in performance.

predicate	m/u	Bert	Roberta	supp.
prefers [p] over [n]	u	0,8	0,841	6662
prefers [n] over [p]	m	0,249	0,457	6634
replaced [n] by [p]	u	0,76	0,761	6700
replaced [p] by [n]	m	0,296	0,494	6777
chose [p] instead of [n]	u	0,812	0,819	6759
chose [n] instead of [p]	m	0,233	0,478	6674
likes [p] but not [n]	u	0,78	0,831	6751
likes [n], but not. [p]	m	0,261	0,474	6687
[p] is better than [n]	u	0,823	0,861	6510
[n] is better than [p]	m	0,193	0,331	6512
[p] is worse than [n]	u	0,287	0,466	6659
[n] is worse than [p]	m	0,742	0,718	6615
has [p] and [n]	u	0,639	0,772	6761
has [n] and [p]	m	0,733	0,771	6537
there is [p] and [n]	u	0,684	0,784	6599
there is [n] and [p]	m	0,694	0,767	6817

Table 10.14: Accuracy score of contextualized language models on the templates with respect to markedness (m/u) of the individual templates. In this analysis, markedness depends on the sequence in which the candidate concepts are mentioned in combination with each predicate.

A possible explanation for this effect may be the positive and negative connotation imposed by the predicate. The word *worse* in close proximity to the correct candidate referent imposes a strong negative association which makes the model less likely to produce it as an answer. In many cases, a candidate that is depicted negatively will not be the correct answer. Five of the other predicates impose a positive connotation on the example they are closest too. The two templates with neutral predicates (*_has and _* and *is _ and _*) show hardly any performance difference. We conclude that word associations from the context may also act as clues for the models. At this point, we cannot determine whether the performance differences are due to discourse structure or connotation imposed by the predicate.

Potential biases in Winogrande

In order to get additional insights into the potential effects of the discourse structure, we attempted an analysis of examples in the Winogrande development set. Specifically, we searched for instances which have the same predicates as our templates and allow for a marked and unmarked discourse structure without changing their meaning. Our search resulted in 14 examples. For each of the 14 examples, we added a marked or unmarked variant (see Appendix). Based on the resulting 28 instances, we cannot see a clear performance difference with respect to markedness (Table 10.15).

This analysis indicates that discourse structure is unlikely to constitute a bias in the Winogrande dataset. Rather, it can be expected that the language models capture conventional

structures in the pre-training process (simply because they are much more common). When not being able to rely on the knowledge required to make the correct decision, they simply rely on the information provided by the discourse structure. The fact that the Winogrande test instances shown in Table 10.15 do not trigger the same model behavior as the templates may be an indication that highly unnatural templates with unexpected discourse structure constitute a high distraction to the language model. To get more insights it could be considered to explore the behavior of language models with respect to discourse structure by means of a larger dataset of minimal pairs of natural sentences.

Markedness	**Bert-ft**	**Roberta-ft**	**Support**
marked	8/14	8/14	14
unmarked	9/14	8/14	14

Table 10.15: Total number of correct predictions of contextualized language models on the filtered and expanded Winogrande development set split based on the marked and unmarked discourse structure.

10.5 Discussion and Conclusion

The goal of this chapter was to assess the semantic knowledge captured by pre-trained and fine-tuned contextualized language models by means of the diagnostic dataset. I have presented an analysis of token probabilities in cloze tasks (Study 1 in Section 10.3) and a template-based challenge constructed around positive and negative examples of semantic properties (Study 2 in Section 10.4). The cloze task assesses to what degree property-specific knowledge is captured by the pre-trained contextualized models. The Winograd-style challenge, in contrast, assesses to what degree models fine-tuned on a common sense reasoning task (in this case the Winogrande task) can reason over semantic properties.

It is possible to draw the following conclusions: Based on the cloze tasks, it is apparent that both pre-trained Bert models capture property-specific knowledge to some degree. Bert-large performed successfully on 13 out of 21 properties. Roberta, in contrast, performed successfully on considerably fewer properties. While Bert outperformed the random baseline for the control property in both task variants, Roberta did not do so in any of the task variants. It should be kept in mind that 'success' was defined as outperforming the best out of 100 tasks with randomized label distributions, which constitutes a high bar. It is, however, surprising Roberta could not beat this bar for gender information, which is highly likely to be encoded in distributional patterns.

It is striking that pre-trained Roberta could only outperform the random baseline for a few properties in the cloze task. A possible reason for this could lie in the difference in pre-training regimes between Bert and Roberta: Both models are trained on masked token prediction. In contrast to Roberta, Bert is also trained on next sentence prediction. It is possible that property information does not occur in immediate proximity to the concept. Rather, it may be mentioned outside of same sentence. If this is indeed the case, Bert can still

capture this information, while Roberta has no way of accessing it. This hypothesis could be explored in further research.

The two variants of the cloze task capture slightly different associations. The analysis of predicted probabilities for either the concept or property indicated that models tend to capture an association from concept to property, but not necessarily the other way around. This could be explained by the tendency of concepts in the dataset to be categorically specific compared to the properties. This difference does not necessarily have consequences for what could be learned on a common sense reasoning task in which both properties and concepts are present in the examples.

The Winograd-style challenge assessed two models fine-tuned on the Winogrande training set. The task contains Winograd instances that should require common sense reasoning. If the fine-tuned models have indeed learned to access common sense knowledge and reason over it, they should perform highly on our template-based challenge. While both models performed above a random (chance-based) baseline, the results remained modest. Roberta achieved 0.94 for the control property **female**, but remained below 0.75 for all other properties. Bert could not go beyond 0.66.

If property-information is encoded in the models, we would expect to see first indications in the pre-trained models tested on masked token prediction. When comparing the performance of the pre-trained to the fine-tuned models for individual properties, it is striking that there is no clear alignment: It is not the case that properties with high results in the token prediction task also lead to high performance on the Winograd-style task. This contrast is particularly stark when considering Roberta: Roberta clearly outperforms Bert in the Winograd-style task, but performed worse than Bert in the cloze tasks. It is particularly surprising that the fine-tuned Roberta model achieved almost perfect performance for the control property in the Winograd-style set-up while the pre-trained Roberta model could not achieve high performance on the cloze task involving the control property. This divergence poses the question of what the models learned during fine-tuning. Could the fine-tuning process indeed foreground information that is not apparent from the pre-training models (as is the case for the property **female**)? What kind of information did the models use to arrive at decisions?

One possibility is that the models learned to exploit superficial features to arrive at the correct answer, such as certain types of discourse structure. To investigate whether the models exploit this information rather than semantic information, we experimented with systematic variations in the templates. The results showed that both models perform considerably higher on unmarked discourse structure variants than on marked ones. This stark performance difference indicates that the models may be much more sensitive to superficial features than to the semantic information expressed in the sentences. Regardless of whether they contain property-information, the signal provided by the discourse structure was stronger than the signal provided by the semantic triggers. Given the strong signal from the discourse structure, it is difficult to tell whether the models fine-tuned on the Winogrande training set learned to reason over the semantic properties and concepts in our diagnostic dataset or whether they only learned to exploit superficial features. Whether they acquired any reasoning abilities about other aspects of common-sense knowledge remains an open question for future research.

The experiments presented in this section have the following limitations: Firstly, the masked token prediction task does not take into account that property information can be expressed in various ways, which may differ with respect to different semantic categories (e.g. **dangerous** may be expressed differently for dangerous animals, criminals, weapons or substances). Secondly, the Winogrande training set may simply not be suitable to fore-ground semantic property knowledge. Rather, it may emphasize aspects that arise from the specific, compositional contexts of the Winogrande sentences. In future work, it could be considered to fine-tune models on sentences that highlight property-specific knowledge.

10.6 Summary

In this chapter, I have taken the first steps towards analyzing semantic property knowledge captured by contextualized language models by means of two tasks. The masked concept prediction task can provide insights into whether property-evidence in text triggers higher probabilities for positive examples of a property than for negative examples in pre-trained language models. Vice-versa, the property-prediction task shows whether positive examples of a property trigger higher probabilities for the property than negative examples. The results showed that pre-trained models follow this behavior for a subset of properties in the diagnostic dataset.

The Winograd-style property task tests whether fine-tuned models trained to perform on the Winogrande challenge can use their knowledge and reasoning abilities to distinguish positive and negative examples of a semantic property. The models achieved above-random, but overall much lower performance on this task than on the Winogrande test set. Additional analysis indicates that the models rely much stronger on discourse structure than on property knowledge. Even if the models capture property knowledge to some extend, the signal is weak and seems to be overpowered by unconventional formulations.

Conclusions

Main Findings

The central focus of this thesis was placed on the investigation of semantic property knowledge in distributional representations of word meaning. Specifically, the research presented in the preceding chapters aimed to answer the following research question:

What aspects of conceptual knowledge are reflected by the co-occurrence patterns captured by large-scale language models?

I have addressed this central research question through three major components: (1) The thesis proposed a model for testing semantic property knowledge in distributional representations (Part II). (2) Based on the model, I have created a diagnostic dataset by means of eliciting semantic judgments from crowd workers (Part III). (3) I have used the diagnostic dataset to design diagnostic experiments for context free and contextualized semantic representations. To complement the experimental results, I have exploited the contrastive nature of the dataset to verify experimental results by means of corpus analysis (Part IV). In this section, I summarize the main findings that arose from the research conducted in the three parts of the thesis. I group them by topic and link them to the three steps I used to operationalize the main research question:

Step 1 : *Create **a model of conceptual knowledge and property expression** for the investigation of language model representations.*

Step 2 : *Capture human conceptual knowledge in a **dataset** suitable for the investigation of language models.*

Step 3 : *Use **interpretability methods** for context-free and contextualized language models to study which aspects of semantic knowledge they represent.*

A Dataset as a Diagnostic Tool

Step 1 The core of the methodological approach taken in this thesis lies in the construction of a dataset as a diagnostic tool. The dataset was constructed in such a way that it follows the methodological challenges of analyzing context-free embedding representations (Chapter 4). The main challenge of diagnosing specific semantic properties in embedding vectors by means of diagnostic classification is the danger of achieving high classification performance on the basis of accidental correlations rather than the fact that the classifiers could identify the target property. Such a scenario is likely to occur if the positive examples of a property can easily be separated from the negative examples of a property on the basis of other salient features. For instance, if all positive examples of the property **red** share a semantic category (e.g. FRUIT:

strawberry, *raspberry*, *cherry*) and all negative examples share a different category (e.g. FURNITURE: *table*, *chair*, *closet*), successful classification performance is not indicative for the target property. To avoid such outcomes, the positive and negative examples of a property have to follow a specific distribution: Ideally, positive examples should represent a diverse set of semantic categories. This was achieved by selecting properties that apply to a diverse set of concepts. In addition, negative examples should be similar to positive examples. This requirement was fulfilled by means of specific selection strategies. If these requirements are fulfilled, successful classification is likely to indicate that the property in question is reflected by the distributional representations.

Step 2 In order to determine whether the dataset resulting from crowd annotations can indeed pose a sufficiently high challenge in diagnostic experiments, I analyzed the dataset with respect to its diagnostic power (Chapter 7). The analysis revealed that the datasets for different semantic properties vary in difficulty; while some properties run risk of being comparatively easy and possibly containing correlations (e.g. the dataset for the property **used_in_cooking**), others can, with relatively high certainty, only be solved if a classifier detects the target information (e.g. **sweet**, **juicy**, **yellow**).

Step 3 The diagnostic experimental set-up for the analysis of context-free embeddings introduced in Chapter 8 shows how the diagnostic dataset in combination with control tasks can be used as a powerful diagnostic tool. The particular distribution of examples in combination with the comparison against control tasks can minimize the change of accidental correlations. The emphasis on challenging examples in the diagnostic dataset (i.e. positive and negative examples of a property that have high semantic similarity, but differ with respect to the target property, such as *duck* and *rabbit* with respect to the property **fly**) allow for a targeted analysis of property representation. The contrastive nature of the diagnostic dataset also allowed for a corpus analysis of property evidence that could be used to verify the results of the diagnostic experiments.

Eliciting Semantic Judgments from the Crowd

Step 2 Another requirement for a diagnostic dataset is that it should contain reliable semantic information. Conceptual knowledge is difficult to capture and can be interpreted differently by different people. This openness for interpretation is a reflection of various linguistic and cognitive phenomena. In order to cover the range of possibilities on the spectrum of clear-cut cases (e.g. **yellow** - *lemon*) to high degrees of ambiguity or vagueness (**fly** - *bat*, **yellow** - *leopard*), I opted for eliciting semantic judgments from crowd annotators.

Eliciting fine-grained semantic judgment from untrained crowd annotators poses a challenge and requires careful task design and monitoring of the annotation process (Chapter 5). In a task for which varying interpretation of annotation units are expected, quality assessment cannot rely on agreement. To establish quality given valid disagreement, I provide a systematic evaluation of alternative quality metrics. I show that simple, coherence-based checks pose an alternative to agreement and can be used to distinguish reliable from unreliable annotations.

The second part of the evaluation presented in Chapter 6 assessed the degree to which crowd annotators were able to make fine-grained semantic distributions. Some property-concept relations, however, seem to be too difficult for untrained crowd annotators given the current task set-up. The two relations expressing a different type of typicality (`typical_of_property`: **red**-*blood* v.s. `typical_of_concept`: **green**-*broccoli*) were not sufficiently distinguished by the annotators. It might be the case this difficulty arises from the task design; both relations are expressed similarly and it is possible that annotators did not spot the difference.

Semantic Properties in Context-Free Representations?

The results obtained from using diagnostic methods, in particular diagnostic classification, are often difficult to interpret. Above random, but clearly not perfect performance of a diagnostic classifier could indicate that property-information is captured by only a subset of examples (valid outcome) or that the classifier identified a spurious correlation that held for some, but not all examples in the dataset, such as a semantic category that happened to correlate with some, but not all positive examples (e.g. BERRY for the property **red**). The latter outcome is misleading. In order to distinguish between valid and misleading outcomes, I used baselines and control tasks against which the diagnostic classifiers could be compared. A good control task 'simulates' a scenario in which a classifier cannot access the target property, but can instead rely on other information that leads to reasonable performance. Performance above the control classifier indicates that the target information was indeed identified successfully.

The diagnostic experiments presented in Chapter 8 show little evidence that property-specific information is systematically represented in context-free distributional representations. The semantic control task in combination with the architecture of the diagnostic dataset provided strong indications that context-free embeddings do not encode information about perceptual properties (colors, temperatures, shapes). This observation was confirmed in an analysis of challenging examples (concept pairs with high semantic similarity that can be distinguished by the target property).

For other properties, the experiments did show at least partial indications that property-information could be encoded (e.g. **square**, **used_in_cooking**, **lay_eggs**, **juicy**). However, the analysis of the diagnostic power of the property datasets for these properties (Chapter 7) indicated that the example distribution for the high performing properties (in particular **square**, **used_in_cooking** and **lay_eggs**) runs risk of containing unwanted correlations. The results of the corpus analysis presented in Chapter 7 cast additional doubt on whether the diagnostic classifiers could indeed identify property-specific evidence.

The results of the error analysis (Chapter 8) in combination with the corpus analysis (Chapter 9) provide first indications that semantic information in the embedding representations captures fine-grained semantic categories rather than specific properties. The classification errors indicate that even well-performing classifiers cannot make distinctions between highly related (or similar) concepts (e.g. **wheels**: *car* vs. *windshield*). However, many examples can be classified correctly on the basis of fine-grained categories (e.g. *luggage* v.s. *passenger* may be distinguishable by means of their semantic category difference). The analysis of property evidence in corpora supports this indication; the most commonly found evidence of semantic

properties is not expressed by means of direct expressions of the target property (e.g. **red**: *red*), but indirectly, through other concepts that share the target property (e.g. **red**: *blood*, *paint*).

Challenges of Analyzing Contextualized Models

Step 3 The diagnostic dataset and methodological framework of this thesis have primarily been designed to analyze context-free embeddings. Contextualized language models have access to the same type of information (i.e. corpus data) as context free models. Thus, many of the core considerations still apply to diagnostic experiments for contextualized models. Nevertheless, their architectures and training regimes differ substantially from context-free embeddings, which poses a number of additional challenges for interpretability experiments. Most importantly, it is not trivial to extract a representation of an individual word, as the model represents words given a particular context in multiple layers of a network.

Rather than using internal representations of the models in diagnostic classification tasks, I opted for two behavioral tasks (introduced in Chapter 10): As an initial approach, I examined pre-trained models by exploiting the contrastive nature of the diagnostic dataset to compare token probabilities in masked token prediction tasks. The tasks tested to what degree the probabilities assigned to tokens given a particular context can distinguish between positive and negative property concept associations (e.g. *The sea is blue.* vs *The apple is blue*). The results indicated that for a subset of properties, systematic differences between positive and negative examples can be detected. However, it remains difficult to assess what (often quite small) differences in token probabilities mean and how such probabilities translate to the reasoning abilities of the models.

To gain a deeper understanding of the potential of the language models to engage in reasoning over properties, I presented an approach in which the diagnostic dataset was used for the automatic generation of a Winograd-style challenge dataset. The task was used to examine whether models pretrained on a large dataset of Winograd sentences (Winogrande) can also perform well on semantic properties. The Winogrande dataset was designed to test the ability of models to engage in common sense reasoning. For all properties, the models achieved performance barely above a random baseline. An examination of potential biases in the template-based dataset revealed that the models seemed to have based their decisions on specific discourse structures rather than property information; the models could not predict the correct answers given an unconventional discourse structure in a high number of instances.

The behavior of the fine-tuned models is not necessarily evidence that language models are not able to reason over semantic properties. Rather, it illustrates the difficulties of examining contextualized models. Fine-tuning may introduce biases that lead to the correct answers for the wrong reason. In the case of the models fine-tuned on Winogrande, it remains difficult to determine whether the behavior of the models is a reflection of such a bias or whether the signal from the discourse structure was simply stronger than the signal that arose from the property-concept combinations.

Modeling the Dynamics of Property Expression in Corpora

Step 1 One of the goals of this thesis was to find out what underlying factors drive the explicit expression of property evidence in texts. Based on theoretical and empirical linguistic research, I have proposed a framework of factors that could impact to what degree conceptual knowledge is made explicit in texts (Chapter 3). The models allows for deriving specific hypotheses about property expressions.

Step 2 To test the hypotheses derived from the model, I have designed and collected a dataset of properties, concepts, and property-concept relations that represent the different linguistic factors that may determine property expression. The crowd annotation task presented in Chapter 5 resulted in a dataset of fine-grained semantic judgments of property-concept pairs that reflect these linguistic factors. The analysis of the dataset (Chapter 7) showed that the individual linguistic factors (e.g. impliedness, affordedness, variability) interact in complex ways, making it difficult to use the annotated dataset for testing individual hypotheses. The corpus analysis presented in Chapter 9 constitutes an attempt to analyze the expression of property evidence with respect to specific linguistic factors on the basis of a selection of property-concept pairs. The analysis is based on few examples and can thus not yield reliable insights.

Future Work

The research presented in this thesis illustrates that the interaction between semantic information, distributional data, and different language model architectures and training regimes is complex and not yet well understood. The methodological considerations and diagnostic experiments highlight the difficulties involved in drawing sound conclusions from different interpretability experiments involving machine learning. The work presented in this thesis has also served to illustrate the importance of careful dataset design and construction that anticipates methodological challenges of interpretability experiments. On the basis of these observations, I propose the following directions for future research:

Data manipulation experiments A possible means of gaining a better understanding of the interaction between distributional models and linguistic co-occurrence patterns could be controlled context manipulation experiments. It could be considered to simulate different types and distributions of property evidence and test how the different evidence constellations affect the outcome of diagnostic experiments. Such experiments could also give insights into how different architectures for context-free models (in particular architectures designed for small data mentioned in Chapter 1 react to linguistic contexts.

Full exploitation of the template-based dataset The template-based approach for generating evaluation data introduced in Chapter 10 could be a promising tool for a closer examination of what signals contextualized models are sensitive to during fine-tuning. While template-based instances have the disadvantage of sounding 'unnatural', they have the advantage of enabling highly controlled experiments. The analysis presented in Chapter 10

showed systematic variations in the templates can be exploited for error analysis. Beyond this, variations in templates can also be exploited to investigate what models learn during fine-tuning. By using different, controlled distributions during fine-tuning, it is possible to explore what kinds of generalizations models tend to make; do they rely on superficial patterns of do they pick up property-specific information?

Extension of the diagnostic dataset As a final point, the diagnostic dataset could be improved on two levels: Firstly, the diagnostic power of the datasets could be increased by strategically adding challenging and informative examples to property datasets with a less challenging example distribution. Secondly, the statements used to express fine-grained relations between properties and concepts could be revised in such a way that crowd annotators can distinguish them with higher reliability. Such a revision may entail a simplification of the original framework of property-concept relations proposed in Chapter 3.

Bibliography

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. arXiv preprint arXiv:2109.06129.

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to winograd schema perturbations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7590–7604.

Astrid van Aggelen, Antske Fokkens, Laura Hollink, and Jacco van Ossenbruggen. 2019. A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, pages 44–54. Linköping University Electronic Press.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. Psychological review, 116(3):463.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. Transactions of the Association of Computational Linguistics, 6:107–119.

Marianna Apidianaki and Aina Garí Soler. 2021. All dolphins are intelligent and some are friendly: Probing bert for nouns' semantic properties and their prototypicality. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 79–94.

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. Human Computation, 1(1).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4):555–596.

Martha Augoustinos and Danielle Every. 2007. The language of "race" and prejudice a discourse of denial, reason, and liberal-practical politics. Journal of Language and Social Psychology, 26(2):123–141.

Martin Barker. 1981. The new racism: conservatives and the ideology of the tribe. Junction Books.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 238–247.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4):673–721.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. Cognitive science, 34(2):222–254.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. arXiv preprint arXiv:2102.12452.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7:49–72.

Arianna Betti and Hein van den Berg. 2014. Modelling the history of ideas. British Journal for the History of Philosophy, 22(4):812–835.

Arianna Betti, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway, and Jelke Bloem. 2020. Expert concept-modeling ground truth construction for word embeddings evaluation in concept-focused domains. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6690–6702.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. Cognitive neuropsychology, 33(3-4):130–174.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Marianna Bolognesi, Christian Burgers, and Tommaso Caselli. 2020. On abstraction: decoupling conceptual concreteness and categorical specificity. Cognitive processing, 21(3):365.

Anna M Borghi and Ferdinand Binkofski. 2014. Words as social tools: An embodied view on abstract concepts, volume 2. Springer.

Anna M Borghi and Nicoletta Caramelli. 2003. Situation bounded conceptual organization in children: From action to spatial relations. Cognitive Development, 18(1):49–60.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 136–145.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. Journal of artificial intelligence research, 49:1–47.

Marc Brysbaert, Paweł Mandera, and Emmanuel Keuleers. 2018. The word frequency effect in word processing: An updated review. Current Directions in Psychological Science, 27(1):45–50.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype bert embeddings for estimating semantic relationships. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 227–244.

Max Coltheart. 1981. The mrc psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4):497–505.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive science, 19(2):233–263.

Mark Davies. 2002. The Corpus of Historical American English (COHA): 400 million words, 1810-2009.

Manuel De Vega, Arthur C Graesser, and Arthur M Glenberg. 2008. Reflecting on the debate. Symbols and embodiment: Debates on meaning and cognition, pages 397–440.

Marco Del Tredici and Núria Bel. 2015. A word-embedding-based sense index for regular polysemy representation. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 70–78.

Steven Derby, Paul Miller, and Barry Devereux. 2019. Feature2vec: Distributional semantic modelling of human property knowledge. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5856–5862.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. Behavior research methods, 46(4):1119–1127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In COLING, pages 3519–3530.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In NetWordS, pages 66–70.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 conference on empirical methods in natural language processing, pages 1136–1145.

A Dumitrache. 2019. Truth in disagreement: Crowdsourcing labeled data for natural language processing.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015. Crowdtruth measures for language ambiguity. In Proc. of LD4IE Workshop, ISWC.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In Sixth AAAI Conference on Human Computation and Crowdsourcing.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2164–2170.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 52–58.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Bias detection, training and commonsense disentanglement in the winograd schema. arXiv preprint arXiv:2104.08161.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? Semantics and Pragmatics, 9:17–1.

Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2003. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 537–544.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8:34–48.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In Proceedings of the 11th International Conference on Computational Semantics, pages 52–57.

Christiane Fellbaum. 2010. Wordnet. In Theory and applications of ontology: computer applications, pages 231–243. Springer.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In Proceedings of the 10th international conference on World Wide Web, pages 406–414.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? In Conference of the Cognitive Science Society (CogSci).

Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. 2017. What can you do with a rock? affordance extraction viaword embeddings. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pages 1039–1045.

James J Gibson. 1954. The visual perception of objective motion and subjective movement. Psychological Review, 61(5):304.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In Proceedings of the NAACL Student Research Workshop, pages 8–15.

Arthur M Glenberg. 1997. What memory is for. Behavioral and brain sciences, 20(1):1–19.

Arthur M Glenberg and David A Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. Journal of memory and language, 43(3):379–401.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 609–614.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In Proceedings of the 2013 workshop on Automated knowledge base construction, pages 25–30.

H Paul Grice. 1975. Logic and conversation.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1489–1501.

WL Hamilton, J Leskovec, and D Jurafsky. 2016b. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2016, pages 2116–2121. NIH Public Access.

Zellig S Harris. 1954. Distributional structure. Word, 10(2-3):146–162.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics-Volume 2, pages 539–545. Association for Computational Linguistics.

Johannes Hellrich and Udo Hahn. 2016a. Bad company—neighborhoods in neural embedding spaces considered harmful. In COLING (16), page 2785–2796.

Johannes Hellrich and Udo Hahn. 2016b. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In LaTeCH 2016—Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities@ ACL, pages 111–117.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: Acquiring new word vectors from tiny data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 304–309. Association for Computational Linguistics.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 22–32.

Aurélie Herbelot and Eva Maria Vecchi. 2016. Many speakers, many worlds. LiLT (Linguistic Issues in Language Technology), 13.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 41(4):665–695.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. convolutional neural networks and incremental parsing, 7(1).

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research, 61:907–926.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657.

Brendan T Johns and Michael N Jones. 2012. Perceptual inference through global lexical similarity. Topics in Cognitive Science, 4(1):103–120.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pages 1637–1648.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval. Computers and the Humanities, 34(1):15–48.

Tassilo Klein and Moin Nabi. 2021. Towards zero-shot commonsense reasoning with self-supervised refinement of language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8737–8743.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4837–4842.

Zoltán Kövecses. 2000. The concept of anger: Universal or culture specific? Psychopathology, 33(4):159–170.

Alicia Krebs, Alessandro Lenci, and Denis Paperno. 2018. Semeval-2018 task 10: Capturing discriminative attributes. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 732–740.

Henry Kučera and Winthrop Nelson Francis. 1967. Computational analysis of present-day American English. Dartmouth.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 166–172.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In International conference on machine learning, pages 2873–2882. PMLR.

G Lakoff and M Johnson. 1980. Metaphors we live by.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2):211.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1403–1414.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. Cognitive Science, 41:677–705.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, et al. 1996. Tsnlp-test suites for natural language processing. arXiv preprint cmp-lg/9607018.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. Italian journal of linguistics, 20(1):1–31.

Alana Lentin. 2005. Replacing 'race', historicizing 'culture'in multiculturalism. Patterns of prejudice, 39(4):379–396.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. Citeseer.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In Proceedings of the eighteenth conference on computational natural language learning, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3:211–225.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pages 13–18.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. Behavior research methods, 37(4):547–559.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. science, 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In HLT-NAACL, volume 13, pages 746–751.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Ann Morning. 2009. Toward a sociology of racial conceptualization for the 21 st century. Social Forces, 87(3):1167–1192.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. Computational Linguistics, 46(2):487–497.

Michael A Omi. 2001. The changing meaning of race. America becoming: Racial trends and their consequences, 1:243–263.

Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021. Challenging distributional models with a conceptual network of philosophical terms. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2511–2522, Online. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. Computational Linguistics, 33(2):161–199.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 823–835.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. Transactions of the Association for Computational Linguistics, 7:677–694.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. PloS one, 10(10):e0137041.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology, 70:153–163.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1778–1789.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. NAACL HLT 2018, page 180.

P Proctor. 1978. Longman Dictionary of Contemporary English. Longman Group, Essex, UK.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 57–61.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Brian Riordan and Michael N Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. Topics in Cognitive Science, 3(2):303–345.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8:842–866.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal lda model integrating textual, cognitive and visual modalities. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1146–1157.

Eleanor Rosch. 1973. Prototype theory.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 726–730.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 975–980. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8732–8740.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 298–307.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1526–1534.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In Proceedings of the 28th International Conference on Computational Linguistics, pages 6863–6870.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 572–582.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1423–1433.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642.

Pia Sommerauer. 2020. Why is penguin more similar to polar bear than to sea gull? analyzing conceptual knowledge in distributional models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 134–142, Online. Association for Computational Linguistics.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 223–233, Florence, Italy. Association for Computational Linguistics.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2019. Towards interpretable, data-derived distributional semantic representations for reasoning: A dataset of properties and concepts. In Wordnet Conference, page 85.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In Thirty-First AAAI Conference on Artificial Intelligence.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In LREC, pages 3679–3686.

Karen Spärk Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7046–7056.

Gerard Steen. 2010. A method for linguistic metaphor identification: From MIP to MIPVU, volume 14. John Benjamins Publishing.

Paul Trichelair, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman, and Fernando Diaz. 2018. On the evaluation of common-sense reasoning in natural language understanding. arXiv preprint arXiv:1811.01778.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In Proceedings of the second workshop on linguistic and neurocognitive resources, pages 1–8.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. Cognitive Science, 44(6):e12844.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Tony Veale. 2013. The agile cliché: using flexible stereotypes as building blocks in the construction of an affective lexicon. In New Trends of Research in Ontologies and Lexical Resources, pages 257–275. Springer.

Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 29.

Gabriella Vigliocco, David P Vinson, William Lewis, and Merrill F Garrett. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. Cognitive psychology, 48(4):422–488.

David P. Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. Behavior Research Methods, 40(1):183–190.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. arXiv preprint arXiv:2004.04877.

Unni Wikan. 1999. Culture: A new concept of race. Social Anthropology, 7(01):57–64.

Howard Winant. 1998. Racism today: Continuity and change in the post-civil rights era. Ethnic and Racial Studies, 21(4):755–766.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 747–763.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics, pages 133–138.

Yadollah Yaghoobzadeh, Katharina Kann, Timothy J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5740–5753.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 236–246.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does bert capture semantics? a closer look at polysemous words. In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 156–162.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 359–361.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 1127–1137.

Usecases: Variation and Change

Detailed overview of hypotheses and outcomes

Table 1 provides an overview of the hypothesized changes in the conceptual system of racism and observed changes in two models.

word1	word2	Hypothesis	Coha-sgns	Ngrams-sgns
racial	cultural	closer	-	-
racial	superior	apart	-	-
racial	inferior	apart	apart	-
racial	blacks	apart	-	-
racial	whites	apart	apart	closer
racial	marriage	apart	-	closer
racial	relationships	apart	-	-
racial	genetics	apart	OOV	OOV
racial	nigger	apart	closer	closer
racial	yankee	apart	-	-
racial	gypsy	apart	-	-
cultural	superior	apart	closer	apart
cultural	inferior	apart	-	apart
cultural	blacks	apart	-	closer
cultural	whites	apart	-	closer
cultural	marriage	apart	-	apart
cultural	relationships	apart	-	-
cultural	genetics	apart	OOV	OOV
cultural	nigger	apart	closer	-
cultural	yankee	apart	-	-
cultural	gypsy	apart	-	-
racial	immigrant	closer	-	apart
racial	foreigner	closer	apart	-
racial	national	closer	-	apart
racial	Turks	closer	OOV	OOV
racial	Arabs	closer	-	-
racial	Jews	closer	apart	-
racial	religious	closer	closer	-
racial	linguistic	closer	-	-
racial	values	closer	apart	closer
racial	attitudes	closer	-	apart
racial	beliefs	closer	-	apart
racial	historic	closer	apart	-
racial	different	closer	-	-

cultural	immigrant	closer	-	-
cultural	foreigner	closer	-	-
cultural	national	closer	closer	-
cultural	Turks	closer	-	-
cultural	Arabs	closer	-	-
cultural	Jews	closer	-	-
cultural	religious	closer	closer	-
cultural	linguistic	closer	-	closer
cultural	values	closer	closer	closer
cultural	attitudes	closer	-	-
cultural	beliefs	closer	-	-
cultural	historic	closer	-	-
cultural	different	closer	closer	-

Table 1: Overview of hypothesized changes and results in of the SGNS model in COHA and the google n-grams. The forms of *racial* and *cultural* have been adapted to match word2 in part of speech and number. *closer* indicates a significant change towards each other and *apart* a significant increase in distance, - means no significant change.

Annotation task

Variability and specification for scalar and binary properties The expression of the variability relation has to be adapted for certain types of properties. For scalar or binary properties (e.g. temperatures, being dangerous), the spectrum for variability is limited to a scale or two options. I use the relation `variability_limited` for the most restricted version: variation between two extremes. The relation `variability_open` represents scenarios in which all values on a scale are possible.

The examples below illustrate the intention behind the relations. A pistol is typically dangerous, but could be harmless if it is not loaded or fake. It is unlikely to find somewhat dangerous pistols. Therefore, the relation `variability_limited` is appropriate. The relation does not apply to instances in which variation on a scale is possible (Negative example 1) or to concepts for which the property is not variabiable (negative example 2; poison is dangerous by definition).

Relation: `variability_limited`

Positive example You can find (a/an) pistol which is dangerous. (A/an) pistol is usually either dangerous or not dangerous. It cannot be a bit more or less dangerous.

Negative example 1: You can find (a/an) casserole which is warm. (A/an) casserole is usually either warm or not warm. It cannot be a bit more or less warm.

Negative example 2: You can find (a/an) poison which is dangerous. (A/an) poison is usually either dangerous or not dangerous. It cannot be a bit more or less dangerous.

Examples for `variability_open` again illustrate the contrast; water can have any temperature on a continuous scale. Firearms, in contrast, are either dangerous or not dangerous.

Relation: `variability_open`

Positive example You can find (a/an) water which is cold. (A/an) water is usually either cold, a bit more or less cold or the opposite of cold.

Negative example: You can find (a/an) firearm which is dangerous. (A/an) firearm is usually either dangerous, a bit more or less dangerous or the opposite of dangerous.

It should be noted that for scalar properties, both variability relations express variation along a limited spectrum. The difference between `variability_limited` and `variability_open` is most likely smaller than for other properties.

Variability and specification for part properties For part properties, it the distinction between a limited and open range of alternatives to the part does not make sense. What is important to know is whether part properties introduce an important distinction between different instances of concepts (similar to black v.s. grey v.s. brown bears) or not (similar to t-shirts of different colors). Therefore, the statements for part properties are formulated as follows:

Relation: `variability_limited`

Positive: You can find (a/an) vehicle which has (a/an) wheels. (A/an) vehicle usually either has an wheels or no wheels. This distinguishes a certain type of vehicle from others.

Negative: ('experiment2', '4') You can find (a/an) automobile which has (a/an) wheels. (A/an) automobile usually either has an wheels or no wheels. This distinguishes a certain type of automobile from others.

Relation: `variability_open`

Positive: You can find (a/an) machine which has (a/an) wheels. (A/an) machine usually either has (a/an) wheels or no wheels but this does not distinguish a certain type of machine from others.

Negative: You can find (a/an) falcon which has (a/an) wings. (A/an) falcon usually either has (a/an) wings or no wings. This distinguishes a certain type of falcon from others.

Crowd Annotation Evaluation

prop	concept	prop_true	uas_true	acc.
green	asparagus	1.00	1.00	✓
green	crocodile	1.00	1.00	✓
made_of_wood	lath	1.00	1.00	✓
used_in_cooking	chickpea	1.00	1.00	✓
black	coffee	1.00	1.00	✓
used_in_cooking	pasta	1.00	1.00	✓
used_in_cooking	onion	1.00	1.00	✓
used_in_cooking	garlic	1.00	1.00	✓
wings	sparrow	1.00	1.00	✓
sweet	loquat	1.00	1.00	✓
blue	sapphire	1.00	1.00	✓
swim	guppy	1.00	1.00	✓
black	currawong	0.88	0.87	✓
made_of_wood	harp	0.88	0.88	✓
wheels	motorbike	0.88	0.88	✓
wheels	coupe	0.88	0.87	✓
round	globe	0.86	0.91	✓
hot	pastry	0.86	0.85	✓
dangerous	gang	0.75	0.77	✓
wheels	jeep	0.75	0.76	✓
used_in_cooking	slicer	0.75	0.76	✓
fly	gallinule	0.75	0.75	✓
green	tank	0.75	0.75	✓
sweet	currant	0.75	0.75	✓
lay_eggs	laridae	0.71	0.72	✓
dangerous	hack	0.71	0.67	✓
hot	toaster	0.67	0.67	✓
black	insect	0.62	0.64	✓
round	barrel	0.62	0.70	✓
red	brick	0.62	0.61	✓

Table 2: 30 randomly chosen examples of the relation `typical_of_concept` ranked by the proportion of positive responses.

prop	concept	prop_true	uas_true	acc.
dangerous	gun	1.00	1.00	✓
warm	heat	1.00	1.00	✓
sweet	candy	1.00	1.00	✓
sweet	honey	1.00	1.00	✓
warm	sun	1.00	1.00	✓
wheels	auto	1.00	1.00	✓
green	lime	1.00	1.00	✓
sweet	cake	0.90	0.90	✓
juicy	raspberry	0.88	0.88	?
made_of_wood	stick	0.86	0.93	✓
hot	firewood	0.86	0.85	✗
juicy	apple	0.78	0.79	✓
roll	cylinder	0.75	0.77	✗
green	spinach	0.75	0.75	?
green	crocodile	0.75	0.75	?
made_of_wood	lumber	0.75	0.74	✓
lay_eggs	merganser	0.75	0.74	✗
warm	coat	0.67	0.67	✓
wheels	buggy	0.67	0.66	✗
used_in_cooking	aubergine	0.67	0.66	✗
yellow	flower	0.67	0.66	?
sweet	raspberry	0.67	0.66	✗
made_of_wood	headstock	0.62	0.64	✗
round	disc	0.62	0.67	✗
swim	grindle	0.62	0.61	✗
green	pear	0.62	0.62	✓
red	ladybird	0.62	0.63	?
juicy	mangosteen	0.62	0.62	✗
juicy	chutney	0.57	0.56	✗
sweet	cider	0.57	0.59	✗

Table 3: 30 randomly chosen examples of the relation `typical_of_property` ranked by the proportion of positive responses.

prop	concept	prop_true	uas_true	acc.
wings	currawong	1.00	1.00	✓
used_in_cooking	pepperoni	1.00	1.00	✓
used_in_cooking	slaw	1.00	1.00	✓
wheels	truck	1.00	1.00	✓
hot	percolator	1.00	1.00	✓
used_in_cooking	oven	1.00	1.00	✓
round	frisbee	1.00	1.00	✓
dangerous	handgun	1.00	1.00	✓
wings	peregrine	1.00	1.00	✓
made_of_wood	wood	0.89	0.88	?
juicy	tomato	0.89	0.89	✓
sweet	huckleberry	0.89	0.89	✓
green	crocodile	0.88	0.87	✓
warm	coffee	0.88	0.87	✓
green	leafs	0.88	0.88	✓
made_of_wood	rafter	0.88	0.88	?
used_in_cooking	steak	0.88	0.87	✓
warm	attire	0.83	0.82	✓
warm	vest	0.75	0.75	✓
hot	roast	0.75	0.73	?
green	lettuce	0.67	0.66	✗
hot	tub	0.67	0.67	✓
juicy	burger	0.67	0.66	?
round	dough	0.67	0.64	✗
sweet	banana	0.62	0.63	✓
square	photocopier	0.62	0.63	✓
green	seaweed	0.62	0.62	?
sweet	plantain	0.60	0.61	✓
green	asparagus	0.57	0.56	✗
hot	dish	0.56	0.56	✓

Table 4: 30 examples of the relation `affording_activity`.

prop	concept	prop_true	uas_true	acc.
lay_eggs	timaliidae	1.00	1.00	✓
fly	nuthatch	1.00	1.00	✓
fly	motacilla	1.00	1.00	✓
roll	wheel	1.00	1.00	✓
fly	avocet	1.00	1.00	✓
fly	stork	1.00	1.00	✓
swim	sunfish	1.00	1.00	✓
lay_eggs	duck	1.00	1.00	✓
swim	scaup	1.00	1.00	✓
swim	shark	1.00	1.00	✓
lay_eggs	pintail	1.00	1.00	✓
swim	anglerfish	1.00	1.00	✓
swim	halfbeak	1.00	1.00	✓
swim	otter	1.00	1.00	✓
fly	ouzel	1.00	1.00	✓
lay_eggs	phalacrocorax	1.00	1.00	✓
lay_eggs	tanager	1.00	1.00	✓
lay_eggs	calidris	0.89	0.89	✓
swim	sheldrake	0.89	0.90	✓
lay_eggs	utahraptor	0.89	0.87	✓
lay_eggs	crocodile	0.88	0.87	✓
swim	shrimp	0.88	0.87	✓
lay_eggs	paridae	0.88	0.87	✓
lay_eggs	strigidae	0.86	0.86	✓
lay_eggs	cotingidae	0.78	0.78	✓
lay_eggs	micropterus	0.78	0.77	✓
fly	pheasant	0.71	0.74	✓
lay_eggs	platypus	0.71	0.72	✓
lay_eggs	creeper	0.62	0.62	✓
swim	goldeneye	0.60	0.60	✓

Table 5: 30 examples of the relation `afforded_usual`.

prop	concept	prop_true	uas_true	acc.
swim	primate	1.00	1.00	✓
roll	candle	1.00	1.00	✓
swim	bloodhound	0.89	0.90	✓
swim	boa	0.89	0.89	✓
roll	rifling	0.88	0.86	?
roll	grip	0.88	0.88	✓
roll	radiator	0.88	0.86	✗
roll	crankshaft	0.86	0.85	✗
roll	paddle	0.78	0.78	✓
swim	panther	0.75	0.74	✓
swim	wolf	0.75	0.76	✓
roll	footrest	0.75	0.73	✗
roll	eyelet	0.75	0.74	✓
swim	basset	0.71	0.71	✓
swim	bear	0.71	0.71	✓
roll	lever	0.71	0.74	✗
swim	pug	0.67	0.66	✓
swim	boar	0.67	0.70	✓
roll	bead	0.62	0.63	✓
swim	pony	0.62	0.63	✓
roll	windscreen	0.62	0.57	✗
swim	pig	0.62	0.64	✓
roll	calliper	0.57	0.61	✗
roll	hammer	0.57	0.57	✗
lay_eggs	tropicbird	0.57	0.56	✗
swim	leopard	0.57	0.57	✓
roll	surfboard	0.57	0.59	✗
swim	glutton	0.56	0.55	✓
swim	mankind	0.56	0.54	✗
roll	noseband	0.56	0.55	✗

Table 6: 30 examples of the relation `afforded_unusual`.

prop	concept	prop_true	uas_true	acc.
used_in_cooking	corn	1.00	1.00	✓
yellow	sun	1.00	1.00	✓
made_of_wood	trestle	1.00	1.00	✓
black	sapsucker	1.00	1.00	✓
green	combretum	0.89	0.89	✓
yellow	daffodil	0.88	0.88	✓
made_of_wood	ladder	0.88	0.88	✓
red	strawberry	0.88	0.88	✓
used_in_cooking	rice	0.88	0.88	✓
red	lentil	0.88	0.89	✓
black	gorilla	0.88	0.88	✓
made_of_wood	fingerboard	0.88	0.87	✓
black	aubergine	0.86	0.86	✓
roll	tub	0.83	0.84	?
made_of_wood	washtub	0.75	0.74	✓
juicy	corn	0.75	0.74	✓
red	marinade	0.71	0.70	✓
roll	saw	0.71	0.72	?
round	pumpkin	0.71	0.74	✓
green	strawberry	0.70	0.70	✓
juicy	dessert	0.64	0.64	✓
dangerous	firearm	0.62	0.62	?
round	beet	0.62	0.62	✓
used_in_cooking	cutlet	0.62	0.63	?
black	sand	0.62	0.60	✓
square	computer	0.62	0.62	✗
sweet	soy	0.57	0.58	✓
sweet	bacca	0.57	0.64	✓
blue	caterpillar	0.57	0.56	✓
juicy	fennel	0.56	0.56	✓

Table 7: 30 examples of the relation `variability_limited`.

prop	concept	prop_true	uas_true	acc.
hot	tortilla	1.00	1.00	✓
black	car	1.00	1.00	✓
square	room	1.00	1.00	✓
blue	umbrella	1.00	1.00	✓
green	dress	1.00	1.00	✓
round	salad	1.00	1.00	?
square	snack	0.88	0.88	✓
hot	bathroom	0.86	0.85	✓
red	clarinet	0.83	0.83	✓
blue	football	0.80	0.78	✓
warm	hat	0.78	0.78	✓
made_of_wood	puppet	0.75	0.76	✓
warm	sock	0.75	0.75	✓
square	clipboard	0.71	0.72	?
juicy	strawberry	0.67	0.67	✓
warm	sandal	0.67	0.67	?
hot	tub	0.67	0.68	✓
warm	linen	0.67	0.67	✓
round	pie	0.67	0.64	✓
dangerous	reduviidae	0.62	0.60	✓
square	washroom	0.62	0.61	✓
hot	firebox	0.62	0.65	✓
made_of_wood	rudder	0.62	0.59	✓
warm	dad	0.60	0.60	✓
warm	anklet	0.60	0.67	?
dangerous	colchicine	0.57	0.58	✓
dangerous	freebooter	0.57	0.57	✓
warm	sword	0.56	0.55	?
square	file	0.56	0.54	?
made_of_wood	shaft	0.56	0.56	✓

Table 8: 30 examples of the relation `variability_open`.

prop	concept	prop_true	uas_true	acc.
square	barrel	1.00	1.00	✓
green	piano	1.00	1.00	✓
warm	lemonade	0.91	0.91	✓
blue	clarinet	0.88	0.87	✓
red	cowpeas	0.86	0.84	✓
green	wasp	0.86	0.86	✓
hot	cooler	0.83	0.84	✓
black	corn	0.83	0.83	✓
black	supermarket	0.78	0.78	✓
black	brick	0.75	0.75	✓
dangerous	naproxen	0.75	0.76	✓
yellow	cherry	0.75	0.76	✓
red	ring	0.75	0.74	✓
dangerous	club	0.71	0.72	✓
hot	can	0.71	0.72	✓
green	aubergine	0.70	0.70	✓
yellow	soursop	0.67	0.67	?
yellow	lagoon	0.67	0.68	✓
round	pasta	0.67	0.67	?
yellow	football	0.67	0.67	✓
wheels	trap	0.62	0.62	✓
red	eye	0.62	0.58	✓
blue	frogfish	0.62	0.62	✓
made_of_wood	windshield	0.62	0.62	✓
hot	soot	0.57	0.58	✓
made_of_wood	flatcar	0.57	0.59	?
black	sheep	0.57	0.56	?
made_of_wood	pedal	0.57	0.57	?
blue	daisy	0.56	0.55	✓
yellow	lavandula	0.56	0.55	✓

Table 9: 30 examples of the relation `rare`.

prop	concept	prop_true	uas_true	acc.
hot	flowerpot	1.00	1.00	✓
round	bottle	1.00	1.00	?
blue	oven	0.90	0.91	✓
yellow	huckleberry	0.90	0.89	✓
used_in_cooking	sickle	0.89	0.88	✓
sweet	salad	0.88	0.88	✓
black	dolphin	0.78	0.74	✓
sweet	fry	0.78	0.77	✓
green	cherry	0.78	0.77	✓
used_in_cooking	spanner	0.75	0.75	✓
wings	scooter	0.75	0.75	✓
sweet	bacon	0.75	0.74	✓
dangerous	penicillin	0.75	0.75	✓
dangerous	pusher	0.71	0.70	✓
red	jeep	0.71	0.67	?
roll	hammer	0.71	0.72	✓
sweet	cucumber	0.70	0.70	✓
blue	aubergine	0.67	0.67	✓
yellow	thundercloud	0.67	0.67	✓
dangerous	cure	0.67	0.66	✓
blue	daisy	0.67	0.66	✓
wheels	chaise	0.62	0.63	✓
swim	bulldog	0.62	0.62	✓
blue	flame	0.62	0.63	✗
square	cds	0.62	0.62	✓
yellow	frog	0.60	0.59	✓
wings	automobile	0.57	0.53	✓
warm	frock	0.57	0.57	?
green	crab	0.56	0.55	✓
blue	grapefruit	0.56	0.55	✓

Table 10: 30 examples of the relation `unusual`.

Distinctiveness of relations

rel1	rel1_with_rel2	rel1_rel2	rel2_with_rel1	rel2
afforded_usual	0.83	0.82	0.99	typical_of_concept
afforded_usual	0.97	0.81	0.83	implied_category
affording_activity	0.90	0.74	0.81	implied_category
implied_category	0.85	0.74	0.85	typical_of_concept
affording_activity	0.91	0.71	0.76	typical_of_concept
rare	0.88	0.61	0.66	unusual
typical_of_concept	0.59	0.58	0.98	typical_of_property
affording_activity	0.62	0.57	0.86	typical_of_property
implied_category	0.56	0.53	0.93	typical_of_property
afforded_usual	0.49	0.48	0.97	typical_of_property
typical_of_concept	0.47	0.32	0.51	variability_limited

implied_category	0.44	0.30	0.49	variability_limited
creative	0.57	0.29	0.37	impossible
affording_activity	0.50	0.29	0.40	variability_limited
afforded_unusual	0.37	0.26	0.46	unusual
afforded_unusual	0.28	0.24	0.62	rare
typical_of_property	0.44	0.21	0.29	variability_limited
typical_of_concept	0.26	0.17	0.33	variability_open
creative	0.43	0.16	0.21	unusual
afforded_unusual	0.28	0.16	0.27	creative
affording_activity	0.24	0.15	0.26	variability_open
afforded_unusual	0.21	0.15	0.39	variability_limited
rare	0.30	0.15	0.22	variability_open
implied_category	0.24	0.15	0.29	variability_open
variability_limited	0.21	0.14	0.28	variability_open
rare	0.34	0.14	0.20	variability_limited
unusual	0.24	0.14	0.24	variability_open
afforded_unusual	0.39	0.13	0.17	implied_category
unusual	0.26	0.13	0.21	variability_limited
typical_of_property	0.25	0.12	0.20	variability_open
creative	0.28	0.12	0.17	rare
impossible	0.21	0.10	0.15	unusual
afforded_usual	0.09	0.08	0.33	variability_limited
afforded_unusual	0.14	0.05	0.07	afforded_usual
afforded_unusual	0.11	0.04	0.05	impossible
creative	0.14	0.04	0.06	variability_limited
creative	0.11	0.03	0.04	variability_open
afforded_unusual	0.07	0.03	0.06	typical_of_property
impossible	0.05	0.03	0.05	rare
afforded_unusual	0.07	0.03	0.04	typical_of_concept
afforded_usual	0.01	0.01	0.03	creative
creative	0.03	0.01	0.01	implied_category
afforded_usual	0.01	0.01	0.02	unusual
implied_category	0.01	0.01	0.02	rare
implied_category	0.01	0.01	0.02	unusual
impossible	0.00	0.00	0.00	variability_limited
creative	0.01	0.00	0.01	typical_of_concept
affording_activity	0.01	0.00	0.01	unusual
affording_activity	0.00	0.00	0.00	creative
creative	0.01	0.00	0.01	typical_of_property
afforded_usual	0.00	0.00	0.02	rare
typical_of_property	0.00	0.00	0.00	unusual
typical_of_concept	0.00	0.00	0.00	unusual
rare	0.00	0.00	0.00	typical_of_concept
implied_category	0.00	0.00	0.00	impossible

Table 13: Analysis of intersections between property-concept pairs annotated with relations. The table shows the proportion of pairs annotated with rel1 that have also been annotated with rel2 and vice-versa. The table also shows the proportion of pairs annotated with rel1 and rel1 out of all pairs annotated with either rel1 or rel2.

prop	concept	prop_true	uas_true	acc.
swim	buzzard	1.00	1.00	✓
lay_eggs	cow	1.00	1.00	✓
used_in_cooking	violin	1.00	1.00	✓
blue	jaguar	1.00	1.00	✓
swim	whiff	0.89	0.90	✓
fly	poacher	0.89	0.88	✓
blue	wasp	0.89	0.89	✓
fly	catfish	0.89	0.89	✓
wings	hearse	0.88	0.87	✓
fly	jalopy	0.88	0.87	✓
fly	rudd	0.88	0.88	✓
wheels	windshield	0.88	0.88	✓
blue	avocado	0.88	0.87	?
lay_eggs	deer	0.88	0.87	✓
wings	sedan	0.80	0.80	✓
blue	ketchup	0.78	0.77	?
lay_eggs	felid	0.75	0.77	✓
black	fir	0.75	0.77	✓
red	hazelnut	0.71	0.73	✓
wheels	gondola	0.71	0.72	?
sweet	wintergreen	0.67	0.66	✓
wheels	freighter	0.62	0.61	?
wings	diver	0.62	0.63	✓
blue	buttercup	0.60	0.61	?
roll	wrench	0.57	0.59	?
roll	hammer	0.57	0.57	?
square	disk	0.57	0.57	✓
swim	sparrow	0.57	0.57	✓
juicy	bulgur	0.56	0.55	✓
made_of_wood	steerer	0.56	0.54	?

Table 11: 30 examples of the relation `impossible`.

prop	concept	prop_true	uas_true	acc.
fly	car	1.00	1.00	✓
wings	automobile	1.00	1.00	✓
fly	boat	0.86	0.85	✓
roll	plastic	0.86	0.90	?
green	raccoon	0.86	0.85	✓
blue	giraffe	0.78	0.78	✓
juicy	chip	0.75	0.76	✓
fly	sharpie	0.75	0.75	✓
made_of_wood	rock	0.71	0.71	✓
fly	roebuck	0.71	0.71	✓
fly	deer	0.71	0.71	✓
hot	vegetable	0.71	0.72	✓
hot	winter	0.70	0.61	✓
fly	lion	0.70	0.67	✓
blue	pit	0.67	0.66	✓
fly	seal	0.62	0.63	✓
wings	dozer	0.62	0.64	✓
dangerous	club	0.62	0.59	✓
round	chicken	0.62	0.62	✓
sweet	bulgur	0.62	0.63	✓
lay_eggs	howler	0.62	0.63	✓
sweet	vinaigrette	0.62	0.62	✓
wings	admiral	0.60	0.58	✓
dangerous	crease	0.60	0.54	?
juicy	emperor	0.60	0.60	✓
round	puree	0.57	0.58	?
juicy	kernel	0.57	0.58	?
made_of_wood	carabiner	0.57	0.56	?
roll	bumper	0.56	0.55	?
fly	steamer	0.56	0.57	✓

Table 12: 30 examples of the relation `creative`.

Dataset analysis

Psycholinguistic features in the property datasets

This section contains an analysis of psycholinguistic features in the diagnostic dataset.

- Figure 1 shows the distribution of concreteness scores.
- Figure 2 shows the distribution of familiarity scores.
- Figure 3 shows the distribution of imageability scores.

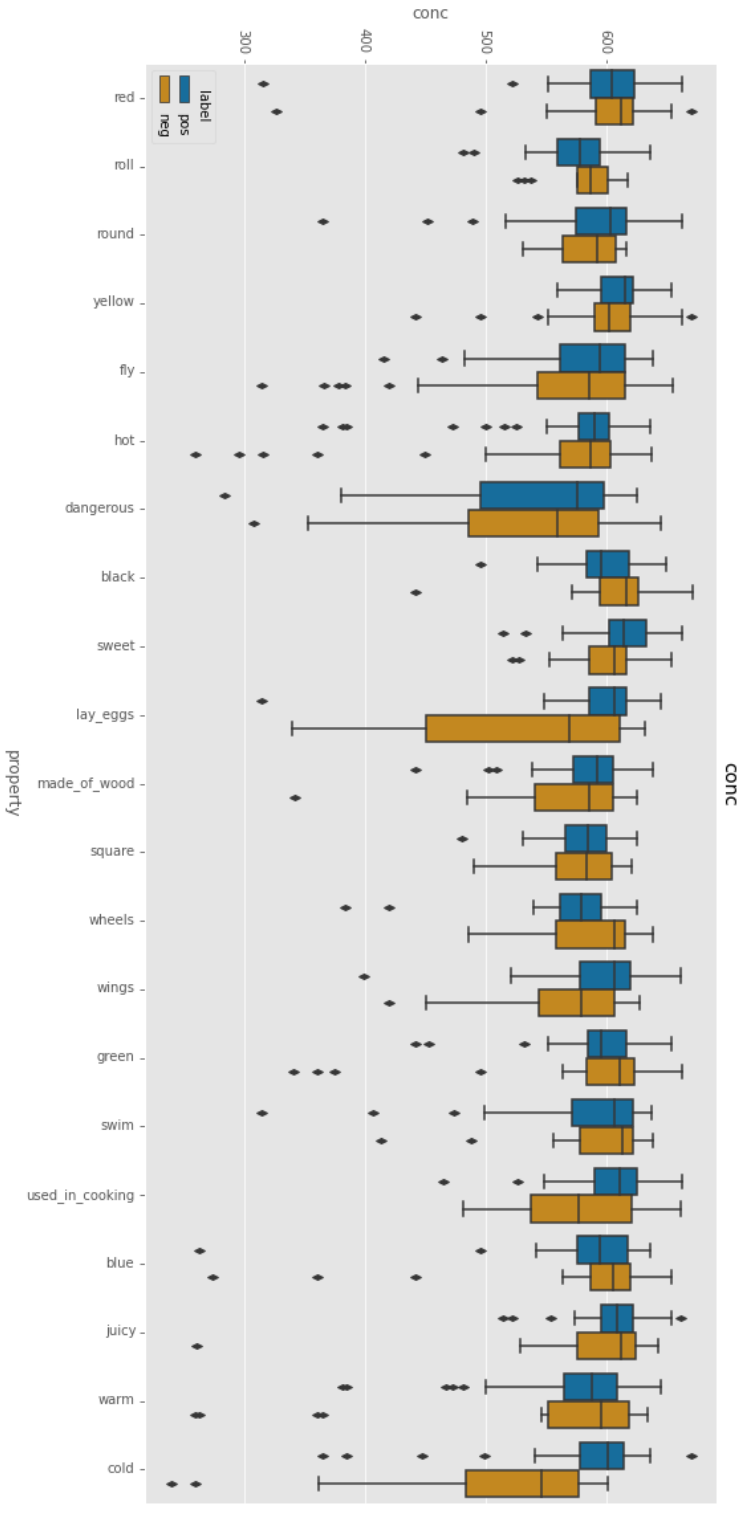


Figure 1: Distribution of concreteness scores.

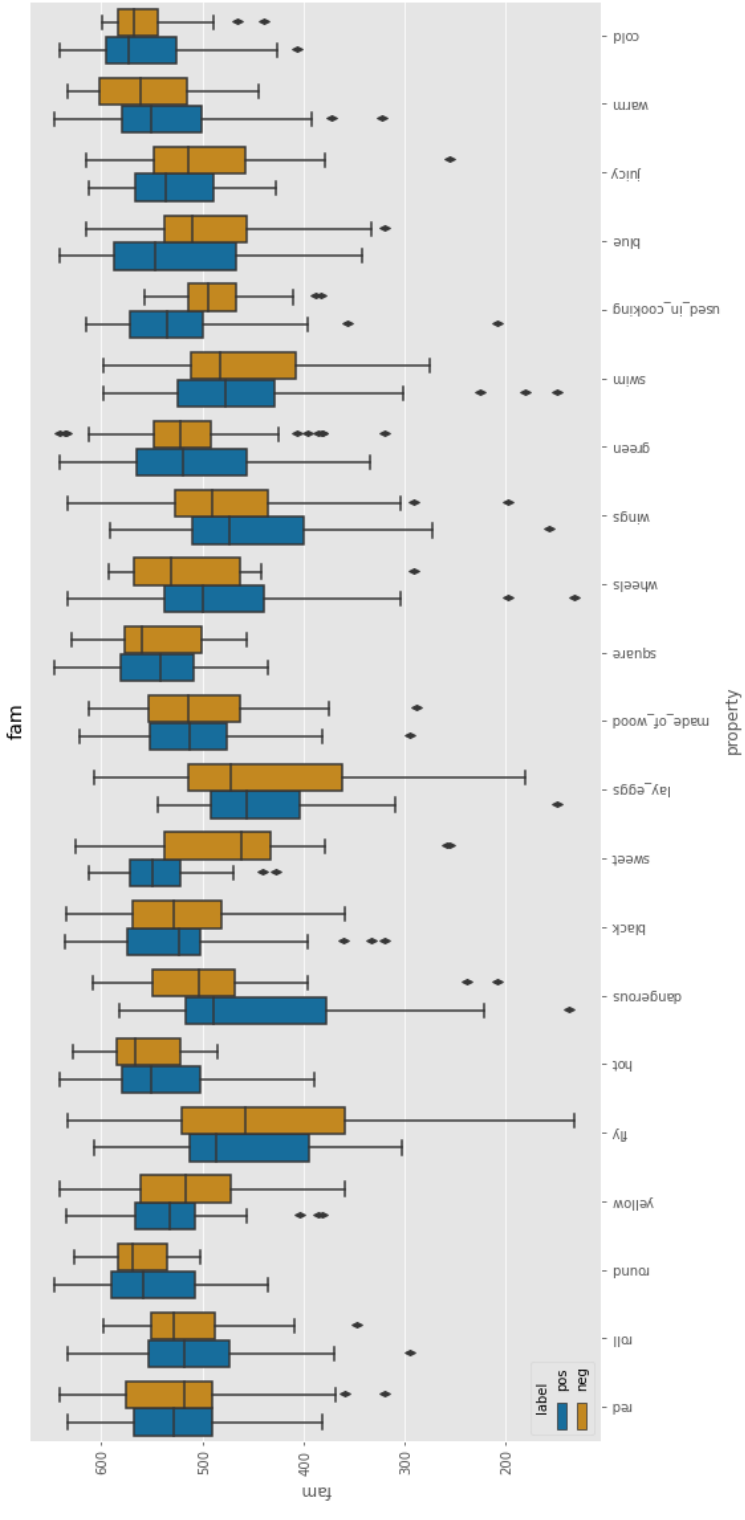


Figure 2: Distribution of familiarity scores.

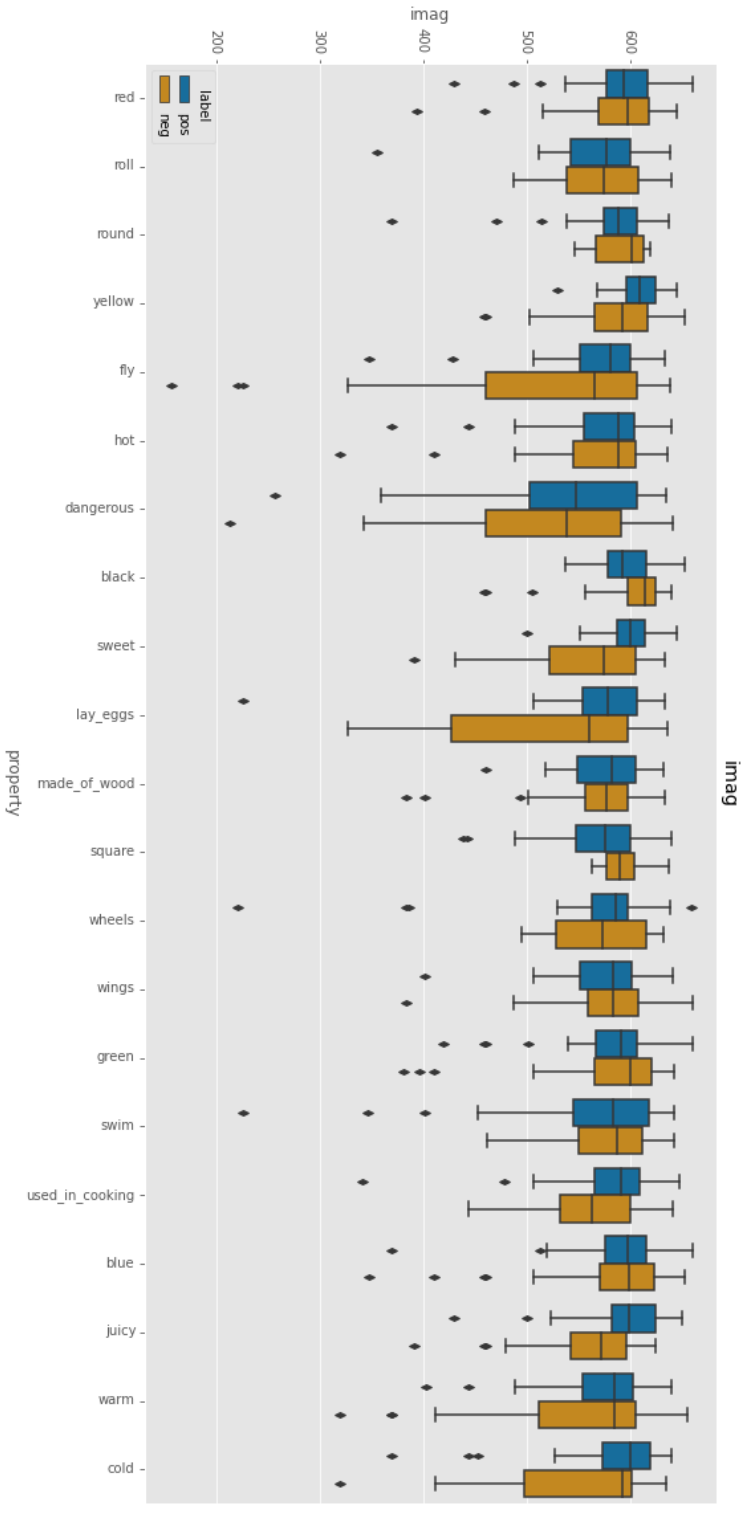


Figure 3: Distribution of imageability scores.

Discourse structure in the Winogrande development set

The full set of marked and unmarked examples extracted from the Winogrande development split can be found in Table 14.

sentence	marked	bert	roberta
Felicia liked wearing glasses more than braces because she could take the _ off after two years.	marked	correct	correct
Felicia liked wearing braces more than glasses because she could take the _ off after two years.	unmarked	correct	correct
Felicia liked wearing glasses more than braces because she could take the _ off every day.	unmarked	incorrect	incorrect
Felicia liked wearing braces more than glasses because she could take the _ off every day.	marked	correct	correct
Pete preferred to use the sheet to the blanket, because the _ was much hotter.	marked	correct	correct
Pete preferred to use the blanket to the sheet, because the _ was much hotter.	unmarked	incorrect	incorrect
Johnny likes fruits more than vegetables in his new keto diet because the _ are saccharine.	unmarked	correct	incorrect
Johnny likes vegetables more than fruits in his new keto diet because the _ are saccharine.	marked	correct	correct
Harold liked to play with dolls more than cars because the _ talked back to him.	unmarked	correct	correct
Harold liked to play with cars more than dolls because the _ talked back to him.	marked	incorrect	incorrect
Mark preferred his drinks in paper cups over styrofoam cups because the _ are strong.	unmarked	correct	correct
Mark preferred his drinks in styrofoam cups over paper cups because the _ are strong.	marked	incorrect	incorrect
She wanted to shop for more clothes and ultimately decided on the velvet dress instead of the denim jacket because the _ was more casual.	marked	correct	correct
She wanted to shop for more clothes and ultimately decided on the velvet jacket instead of the denim dress because the _ was more casual.	unmarked	incorrect	incorrect
The student liked writing their signature with a pen instead of a pencil, because the _ showed up lighter.	marked	correct	incorrect
The student liked writing their signature with a pencil instead of a pen, because the _ showed up lighter.	unmarked	correct	correct

The student liked writing their signature with a pen instead of a pencil, because the _ showed up darker.	unmarked	incorrect	incorrect
The student liked writing their signature with a pencil instead of a pen, because the _ showed up darker.	marked	correct	correct
Aaron wanted to go the gym but the others wanted to go to the park because the _ did require membership.	marked	incorrect	correct
Aaron wanted to go the park but the others wanted to go to the gym because the _ did require membership.	unmarked	correct	correct
During the summer, I like visiting the zoo more than the aquarium because the _ is inside.	marked	incorrect	incorrect
During the summer, I like visiting the aquarium more than the zoo because the _ is inside.	unmarked	correct	correct
The musician liked playing at the auditorium more than at the park because he sounded quieter at the _ .	marked	correct	correct
The musician liked playing at the park more than at the auditorium because he sounded quieter at the _ .	unmarked	incorrect	incorrect
The musician liked playing at the auditorium more than at the park because he sounded louder at the _ .	unmarked	correct	correct
The musician liked playing at the park more than at the auditorium because he sounded louder at the _ .	marked	incorrect	incorrect
She chose the black car over the green car, because the _ has more brighter color.	marked	incorrect	incorrect
She chose the green car over the black car, because the _ has more brighter color.	unmarked	correct	correct

Table 14: Overview of marked and unmarked Winogrande examples extracted from the development split.

Summary

When we communicate with each other, a large chunk of what we express is conveyed by the words we use. Computational models of language often rely on data-derived vector representations of words. Such representations are based on large text corpora and capture the many different contexts in which a word is used. For instance, the word *lemon* is represented by all instances of the word *lemon* in a large text corpus. It can be argued that the meaning of a word can be best characterized in terms of how the word is used. However, it is difficult to determine what aspects of word meaning the text corpora underlying the representations contain. Furthermore, it is unclear how the computational models used to create the word representations react to different usage examples in the data.

Existing computational models of language perform well in some scenarios, but also make silly mistakes humans would never make. Their successes and failures are likely to be caused by what they know (or do not know) about the meaning of words. However, it is unclear what aspects of word meaning data-derived representations capture and what they do not capture. Do computational models know that lemons are yellow and round and have a sour taste?

Usage-based word representations capture the meaning of words in terms of similarities between usage patterns. Words that appear in similar linguistic contexts receive similar vectors. Beyond general semantic similarity, it is very difficult to interpret such word representations. Essentially, they constitute lists of numbers that only become meaningful when put into relation to one another. For example, it is likely that the words *lemon* and *orange* have more similar representations than the words *lemon* and *computer*. It is, however, not possible to see why *lemon* and *orange* are similar. This is particularly problematic in cases where the similarities reflected by the computational representations do not correspond to human judgments of word similarities. This lack of transparency and interpretability has been investigated in several areas of research outlined in Chapter 1. Chapter 2 presents two use-cases that illustrate the limitations of such representations and the need for a better understanding of what they represent when used to study words in specific texts or collections of texts.

The main goal of this thesis is to shed light on the semantic representations of words, which form the basis of many current computational models of language. What type of semantic information tends to be expressed well through usage patterns in text corpora? What aspects of semantic information tend to be absent from such corpora? Linguistic theories and observations provide some indications about what we could expect from text corpora. Chapter 3 brings together several theoretical approaches and observations and introduces a framework of hypotheses about what we can expect from usage-based word representations.

To answer these questions, I draw on methods used to interpret representation created by neural networks. Neural networks can learn to capture complex correlations, but it is not

easy to get insights into how they capture them. Methods that analyze such networks are therefore referred to as 'diagnostic' methods. These methods are still relatively novel and struggle with a number of challenges, in particular when used for the interpretation of word representations. Such methods also require diagnostic datasets to study the representations of specific words. Such methods require a set of informative example words. For instance, we can test whether a model knows that some birds (e.g. *seagulls* and *pigeons*) can fly while others cannot (e.g. *penguins* and *ostriches*). The instances in such a dataset should not be too obvious. Diagnostic methods are informative if the distinctions between words are challenging. Chapter 4 considers the methodological challenges of diagnostic methods and presents the design of such a challenging diagnostic dataset.

The three chapters making up Part III of the thesis focus on the collection (Chapter 5), evaluation (Chapter 6), and analysis (Chapter 7) of the diagnostic dataset. Crowd annotators judged a selection of semantic properties (i.e. aspects of semantic information, such as **being able to fly** and **having a yellow color**) and concepts (e.g. *lemon*, *seagull*, *penguin*). This resulted in a dataset of 21 semantic properties. Each property has positive examples (e.g. the property **yellow** has examples such as *lemon* and *daffodil*) and negative examples (e.g. the words *orange* and *bluebell* do not carry the information **yellow**). The positive and negative examples have been selected in such a way that they are difficult for a model to distinguish.

Several challenges are involved in the collection of fine-grained semantic judgments from untrained linguistic annotators. Words are often ambiguous or vague and people have different interpretations of whether a word is indeed associated with a specific property or not (e.g. Would you describe a leopard as yellow?). At the same time, it is necessary to establish whether crowd annotators deliver reliable judgments (instead of just randomly selecting answers). This means that a substantial degree of disagreement between annotators can be expected. Traditionally, this has been assessed by checking whether annotators agree with each other.Chapter 6 proposes an alternative evaluation of crowd annotations that does not rely on agreement. The method checks whether annotators deliver coherent responses. To complement this evaluation, Chapter 7 provides an analysis of all 21 property datasets in terms of their suitability for diagnostic experiments (Chapter 7).

Part IV presents experiments and analyses of language model representations on the basis of the diagnostic dataset. A major focus of the experimental approaches in this part is the interpretation of results in diagnostic set-ups. Representations of word meaning are represented in complex high-dimensional vector spaces. When testing whether word representations can be distinguished with respect to specific information (e.g. **yellow**: *lemon* v.s. *bluebell*), it is difficult to determine whether the word representations have indeed been distinguished on the basis of the target information or on the basis of other, interfering factors. A main contribution of this part is the design of various baselines and control tasks that help to determine whether a model does indeed capture a specific semantic property. The results of the experiments together with an analysis of corpus data indicate that word representations are unlikely to capture specific semantic properties (e.g. the fact that lemons are yellow). Rather, they seem to reflect information about semantic categories (e.g. the fact that lemons are a type of citrus fruit). These findings confirm tendencies that have already become apparent in previous research from the perspective of model interpretability.