FACETS OF VERB MEANING:
A DISTRIBUTIONAL INVESTIGATION OF GERMAN VERBS

DISSERTATION
ZUR ERLANGUNG DES AKADEMISCHEN GRADES
DOCTOR PHILOSOPHIAE (DR. PHIL.)

eingereicht an
der Sprach- und literaturwissenschaftliche Fakultät
der Humboldt-Universität zu Berlin

von
William Roberts

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr. Julia von Blumenthal

Dekan der Sprach- und literaturwissenschaftliche Fakultät
Prof. Dr. Stefan Kipf

Gutachterinnen/Gutachter: 1. Prof. Dr. Markus Egg
2. PD Dr. Evangelia Kordoni

Tag der mündlichen Prüfung: 6 Dezember 2022

The future's not what it used to be.

— Yogi Berra

Dedicated to my wife and parents.

ABSTRACT

This dissertation provides an empirical investigation of German verbs conducted on the basis of statistical descriptions acquired from a large corpus of German text. In a brief overview of the linguistic theory pertaining to the lexical semantics of verbs, I outline the idea that verb meaning is composed of argument structure (the number and types of arguments that co-occur with a verb) and aspectual structure (properties describing the temporal progression of an event referenced by the verb). I then produce statistical descriptions of verbs according to these two distinct facets of meaning: In particular, I examine verbal subcategorisation, selectional preferences, and aspectual type. All three of these modelling strategies are evaluated on a common task, automatic verb classification. I demonstrate that automatically acquired features capturing verbal lexical aspect are beneficial for an application that concerns argument structure, namely semantic role labelling. Furthermore, I demonstrate that features capturing verbal argument structure perform well on the task of classifying a verb for its aspectual type. These findings suggest that these two facets of verb meaning are related in an underlying way.

ZUSAMMENFASSUNG

Diese Dissertation bietet eine empirische Untersuchung deutscher Verben auf der Grundlage statistischer Beschreibungen, die aus einem großen deutschen Textkorpus gewonnen wurden. In einem kurzen Überblick über linguistische Theorien zur lexikalischen Semantik von Verben skizziere ich die Idee, dass die Verbbedeutung wesentlich von seiner Argumentstruktur (der Anzahl und Art der Argumente, die zusammen mit dem Verb auftreten) und seiner Aspektstruktur (Eigenschaften, die den zeitlichen Ablauf des vom Verb denotierten Ereignisses bestimmen) abhängt. Anschließend erstelle ich statistische Beschreibungen von Verben, die auf diesen beiden unterschiedlichen Bedeutungsfacetten basieren. Insbesondere untersuche ich verbale Subkategorisierung, Selektionspräferenzen und Aspekt. Alle diese Modellierungsstrategien werden anhand einer gemeinsamen Aufgabe, der Verbklassifikation, bewertet. Ich zeige, dass im Rahmen von maschinellem Lernen erworbene Merkmale, die verbale lexikalische Aspekte erfassen, für eine Anwendung von Vorteil sind, die Argumentstrukturen betrifft, nämlich semantische Rollenkennzeichnung. Darüber hinaus zeige ich, dass Merkmale, die die verbale Argumentstruktur erfassen, bei der Aufgabe, ein Verb nach seiner Aspektklasse zu klassifizieren, gut funktionieren. Diese Ergebnisse bestätigen, dass diese beiden Facetten der Verbbedeutung auf grundsätzliche Weise zusammenhängen.

PUBLICATIONS

The following publications are connected to this dissertation:

Egg, Markus, Helena Prepens and Will Roberts (2019). 'Annotation and automatic classification of aspectual categories'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3335–3341.

Roberts, Will and Markus Egg (2014). 'A comparison of selectional preference models for automatic verb classification'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 511–522.

Roberts, Will, Markus Egg and Valia Kordoni (2014). 'Subcategorisation acquisition from raw text for a free word-order language'. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014))*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 298–307.

ACKNOWLEDGMENTS

What I owe to my parents Anne and Ken Roberts cannot be put into words, but I want to express my profound admiration for their unwavering love and support, for their infectious interest in languages, computers, and academic studies, and for their dedication to the virtues of empiricism, rationality, and hard work. My riches also include my accomplished, witty, wonderful sisters Mary and Virginia Roberts. I am humbled and blessed to have been so gallantly welcomed into a new family by Aghmorad and Mohammad Saeed Hajimoradlou, and Fatemeh Sayadchi; kheyli moteshakkeram.

I would like to thank early counsellors and mentors Nick Graham, Robin Dawes, Brian Butler, and the late Roger Browse; Saarbrücker amigos Edi Bartelmes, and Claire and Sascha Auda; Portuguese friends Eugenia Mussa and João Ricardo Silva; Parisian pals Margot Colinet, Lynn Hoendervanger, Benno Rem, and Plamen Turkedjiev; and Berliner buddies Amaya Collantes, Micha Nordheim, Rafael Carrión, Lorena Valdes, Svan Klafack, Yvonne Behnke, Uta Giegengack, and Matthias Franz.

I am grateful to Andreas Eisele and Antonio Branco, who took me on board and taught me many useful and important skills. I have also learned much from Kostadin Cholakov, Ben Roth, Max Jakob, Verena Stein, Danielle Ben-Gera, Dmitrijs Milajevs, Aida Nematzadeh, Martin Popel, and Jan Hajič. At the Humboldt University, fond thanks go to Susann Röthke and Lothar Peter, as well as to my fellow doctoral students Sophia Döring, Nathalie Scherf, and Beate Bergmann. Valia Kordoni taught me so many things, perhaps most importantly how to write and present academic papers. Thanks to Sabine Schulte im Walde for her input on how to evaluate subcategorisation frame acquisition for German, and to Helena Prepens for her painstaking work annotating the lexical aspect of German verbs.

I have been materially and spiritually buttressed by an entire regiment of fine folks. My thanks to Dave Kellam and Alexander MacDonald, who officiated at my wedding; to Axel Hartfiel and Serpil Çelik for making enquiries and organising a much-needed COVID timeout; and to Steven Klett and Hedvig Johansson for providing shelter and home furnishings. I'm lucky to know the artful Jean-Christophe Meillaud, who has magicked so many things into existence for me, as well as the awesome Ozel Annamanthadoo, who years ago inspired and pushed me to embark on this whole great adventure. I've maintained vestiges of sanity in part thanks to my birthday twin Monica Pessoa and lifestyle guru Christian Krause. Thank you to Constantin König, for digging up great music; to Raveesh Meena, who has been

my role model many times; and to Meggie Bouchard Bergevin, for upping my skills in the kitchen and in the office. Most of all, my thanks go to the wise and beautiful love of my life, Sara Hadji Moradlou, for keeping her eyes on the prize, feeding and nurturing, following and leading, comforting, cajoling, reading, discussing, laughing and listening.

Last but not least, I am thankful to my supervisor Markus Egg, whip-smart, eloquent, charming, warm and funny, combining unimpeachable professionalism and saintly patience.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BHT	binomial hypothesis test
CFG	context free grammar
EM	Expectation-Maximisation
HPSG	Head-Driven Phrase Structure Grammar
IE	information extraction
IR	information retrieval
LAS	labelled attachment score
LDA	latent Dirichlet allocation
MT	machine translation
NER	named entity recognition
NLP	natural language processing
NP	noun phrase
PCA	principal components analysis
PCFG	probabilistic context free grammar
POS	part of speech
PP	prepositional phrase
PTB	Penn TreeBank
SCF	subcategorisation frame
SMT	statistical machine translation
SP	selectional preferences
SRL	semantic role labelling
SVD	singular value decomposition
SVM	support vector machine
STTS	Stuttgart-Tübingen tag set
VP	verb phrase
WSD	word sense disambiguation
WSJ	Wall Street Journal
WSM	word space model

Part I

INTRODUCTION

1

INTRODUCTION

This thesis, submitted for a degree in general linguistics, focuses on the lexical semantics of verbs. I will argue that verb meaning encompasses two broad kinds of information: argument structure, and aspectual structure. The thesis will then investigate both of these facets of verb meaning using a distributional approach, by constructing statistical models from a large corpus of text.

The approach followed here is certainly more empirical than theoretical, and is intended to describe the semantics of verbs in a way that can be useful for solving natural language processing (NLP) problems; to this end, I investigate verbs using data representative of general language use that reflects a wide spectrum of topics, registers, and target audiences, observing the semantic behaviour of verbs 'in the wild'. The availability of enormous amounts of raw text from the Web has grown dramatically in the last two decades; harvesting information about verb meaning from large quantities of text in this fashion recommends the use of statistics, and the work reported here often makes use of probabilistic descriptions of language phenomena. Some familiarity with statistics and machine learning techniques is assumed.

The language treated here is German, although much of the theory, and many of the methods and lessons learned would be applicable to other languages, and I will often make use of English examples to illustrate various points. From the point of view of NLP, there are not as many linguistic resources for German as there are for English. The resources that are available, however, are of high quality; I hope that, in the course of the work described here, I have been able to contribute something.

Lexical semantics, the study of word meaning, is a branch of linguistics that has been afforded greater attention in the last few decades, a move away from the older view that 'the lexicon is really an appendix of the grammar, a list of basic irregularities' (Bloomfield, 1933, p 274). An example of this increased emphasis is the waxing role that lexicalisation plays in modern formal theories of grammar, whereby semantic components of words assume a central role in explicating the syntax of a language. Lexicalised grammar frameworks such as Lexical Functional Grammar (Kaplan and Bresnan, 1982), Head-Driven Phrase Structure Grammar (HPSG, Pollard and Sag, 1994), and Lexicalised Tree-Adjoining Grammar (Schabes, Abeillé and Joshi, 1988) became the dominant form of computational deep linguistic analysis in the 1980s and 1990s. In the last few years, distributional semantic

methods have been widely employed, following the *distributional hypothesis*, which postulates that words that occur in similar contexts tend to have similar meanings (Harris, 1954), or, more pithily, 'You shall know a word by the company it keeps.' (Firth, 1957, p 11) This focus on words and their distribution in text has brought with it, variously, new techniques in lexical representation, such as word space models, as well as the kind of modelling of the syntactic behaviour of words that underlies lexicalised probabilistic context free grammar (PCFG) parsers. More recently, NLP research has turned to *word embeddings*, another type of distributional semantic representation of lexical meaning, which are often used in conjunction with neural or *deep learning* techniques.

1.1 WHY VERBS?

Verbs are one of the fundamental and most common parts of speech; at the same time, their complex syntactic behaviour that interacts with their lexical semantic content makes them arguably the 'the lexical category that is most difficult to study' (Fellbaum, 1990, p 40). As linguistic theories commonly make the assumption that semantic structure is 'built around a central verb' (Chafe, 1970, p 10), it seems intuitive that having a high-quality and high-fidelity lexical description of verbs would be favourable to all manner of language processing.

In German, the verb is present in all major sentence types; under the view assumed in this work, the main verb of a sentence to a great extent determines both which arguments will co-occur with it, and, ultimately, the syntactic structure of the entire sentence. For example, if the main verb of a sentence is 'kill', the verb will tend to select for an active sentient subject (Agent) and a passive animate direct object (Patient); furthermore, these roles will tend be realised by words drawn from particular lexical fields. The aspectual properties of the verb will further influence and constrain what kinds of sentential modification are possible. As I will outline in the next chapter, the information that guides these processes in the production of language is lexical in nature. I will also motivate the view that the syntax of verbs follows from their semantics; consequently, careful observation of the structures employed in actual language usage should be able to shed light on the meanings of verbs.

Data-driven acquisition of lexical semantics is important to the future development of NLP. After all, automatic construction of lexical resources is preferable to manual construction for NLP applications, since describing words by hand is time-consuming, and human annotators may omit or mislabel lexical items (Boguraev and Briscoe, 1987). The benefits may not be limited to a single part of speech, because work on verbs may be extensible to other kinds of predicates, such as deverbal nouns (Iordachioaia, Plas and Jagfeld, 2016).

1.2 OUTLOOK

Chapter 2 gives a high-level overview of the linguistic theory relevant to verb semantics, introducing the dichotomy of argument structure and aspectual structure. Argument structure concerns what kinds of subjects and objects a given verb likes to co-occur with; this is a property of the syntactic construction of a sentence, information which can be drawn from treebanks, or – more circuitously – from raw text by using the output of an automatic parser. A syntactic analysis is essential for the investigation of argument structure in German. While in English, the relatively fixed word order makes the identification of grammatical relations possible using a simple chunk parser (Manning, 1993), German has more free word order, and subjects and objects do not always appear in predicable places in the sentence. Case (accusative, dative, genitive) can help to identify grammatical relations, but it is neither unambiguous, nor universally marked. Given the syntactic analysis of a sentence, the free word order of the German language can be easily handled, and the argument structure of the sentence's main verb can be directly observed. In contrast to argument structure, the aspectual structure of a verb has to do with the 'shape' of events in time, and insights into verb aspect can be gleaned only indirectly from grammatical relationships.

Chapter 3 reviews linguistic resources and tools relevant to this dissertation and surveys what is available for German. In this chapter I also develop the textual corpus used as a primary data source for the rest of the thesis. Chapter 4 surveys prior work and related research of direct relevance to the later chapters.

Chapters 5 and 6 address argument structure, taking on, respectively, verbal subcategorisation, and selectional preferences. This work is located at the *syntax-semantics interface*, the junction where a verb's semantic behaviour can be at least partially captured by its syntactic behaviour.

Since the number and types of arguments to a verb are reflections of the verb's underlying meaning, we can begin to estimate which syntactic constructions (alternations) the verb enters into by counting the different ways a particular verb can be instantiated. This in turn provides insights into the verb's meaning, and allows the verb to be grouped with others that exhibit similar behaviour. To this end, chapter 5 reports on work to construct a system for automatically finding the subcategorisation frame (SCF) of a verb in context. This chapter introduces an evaluation paradigm, automatic verb clustering, which measures the quality of a statistical description of a verb by judging how well this description can be used to group together verbs that have related meanings. I then carry this evaluation paradigm forward, using it repeatedly as I develop increasingly detailed representations of verb meaning.

Verbal predicates are realised in syntactic constructions that link together collections of entities denoted by the verb's arguments. The argument structure of a verb will tend to place restrictions on the kinds of arguments that can make an appearance. For example, the verb 'eat' may co-occur with a subject, the eater, and a direct object, the eaten. Tendentially, the subject is *animate* (alive) and perhaps human, and the object is *concrete* (physical and solid) and likely drawn from a lexical field covering kinds of foodstuffs. Thus, we predict that example 1.1 is an unremarkable use of the verb, whereas example 1.2 is semantically odd (indicated here with the question mark):

(1.1) Lina is eating a carrot.

(1.2) ? The little rug is eating geography lessons.

Chapter 6 reports on work to characterise which arguments occur with which verbs, and how they are linked. Because the lexical choice of which arguments a verb will take is directed by the verb's meaning, statistics on verb-argument co-occurrence can also be used to group verbs together into classes that signify similar situations. Thus, this work is evaluated using the automatic verb clustering paradigm established in the previous chapter.

Chapters 7 and 8 present work to delve into the aspectual structure of German verbs. In chapter 7, I first explore heuristic indicators that can shed light on verbal aspect using only syntactic structure; when this approach runs out of steam, I construct a manually-annotated resource that directly labels German verbs with features of their aspectual structure. Chapter 8 then explores how the information contained in this resource can be used for a variety of NLP tasks. I present two applications. The first is an experiment on semantic role labelling, a common NLP task. The second application, which should be quite familiar at this point, is the automatic verb clustering experiment. Here I operate under the hypothesis that verbs belonging to the same verb class will tend to share aspectual structure as well their argument structure. My investigations show that features capturing the aspectual behaviour of verbs are helpful for semantic role labelling, an undertaking that overtly concerns the argument structure of a verb. Furthermore, a final study demonstrates that a statistical description of a verb on the argument structure level can be used to determine the aspectual type of that verb. These results suggest that the two facets of verb meaning overlap each other, and are not fully independent.

Chapter 9 concludes with discussion and final thoughts.

2

THEORETICAL BACKGROUND

This chapter provides a foundation for the rest of the thesis by presenting a brief overview of linguistic theory as it relates to the category of verb. Section 2.1 sketches a definition of the category, and section 2.2 presents a short overview of verb semantics. Sections 2.3 and 2.4 then dive deeper, outlining argument and aspectual structure, the two main kinds of lexically-specified verb behaviour. Section 2.5 discusses other verb-related matters that I will not treat here.

2.1 WHAT IS A VERB?

The verb is a major part of speech of language. This grammatical category can be identified on various levels of analysis.

Morphologically, categories can be identified using *inflectional* or *derivational* evidence. For instance, both English and German allow ready identification of verbs from their inflectional morphology. English verbs have several inflected forms in addition to their uninflected *base form*:

1. past-tense suffix *+(e)d*
2. third person singular present tense suffix *+s*
3. present/imperfective/progressive participle suffix *+ing*

This delineation of the category of verb is a good first approximation, but it is complicated by irregularity in inflectional morphology; many verbs have irregular past or perfective forms, such as 'cut' (at once the present, simple past, and past participle). Furthermore, verbs can be synthesised from of other parts of speech by processes such as zero-derivation (Bauer and Valera Hernández, 2005), whereby, e. g., a name or a noun becomes a verb, such as 'to boycott', coined for the unlucky Charles Cunningham Boycott (1832–1897).

Syntactically, verbs can be identified *distributionally*; that is, they can occupy a variety of particular positions inside of phrases or sentences. In English, for example, Radford (1997, p 33) points out that only a verb in its *infinitival* or uninflected base form can appear in this position:

(2.1) They/it can _____.

This rule is illustrated by the following sentences (ungrammaticality is indicated here with an asterisk):

(2.2) (a) They can stay/leave/hide/die/start/cry (verb)

(b) * They can gorgeous (adjective)/happily (adverb)/down (preposition)/door (noun)

Verbs are often carriers of *finiteness*, that essential property of independent clauses. Finite verbs are marked for tense, subject agreement, and mood (although English subjunctives and imperatives lack subject agreement, Quirk et al., 1985, p 150).

Moreover, a verb is also a word which takes a subject or object. The word 'reads' in example 2.3 can be seen to be a verb because it is inflected to agree with the subject ('Claire') in number (i. e., it is bound to the singular present tense *+s* morpheme from the list above):

(2.3) Claire reads the newspaper.

(2.4) Claire and Sascha read the newspaper.

The inflection changes if the subject is altered to have plural number, as in example 2.4.

This proclivity of verbs to associate with other words in the same sentence will be revisited in section 2.3.

2.2 WHAT DO VERBS MEAN?

One way to introduce a discussion of the meaning of grammatical categories such as verbs is through the lens of *reference*, the ability of linguistic expressions to pick out parts of the world. For instance, proper names can be used to identify real existent individuals or locations:

(2.5) **Justin Bieber** is a singer and songwriter.

(2.6) **Toronto** is the capital of the Canadian province of Ontario.

In example 2.6, the proper name 'Toronto' is being used to refer to the city. Noun phrases (NPs) can also be used to refer, although they do not necessarily do so in all cases:

(2.7) **An obelisk** stands in the Hippodrome of Constantinople in the modern city of Istanbul, Turkey.

(2.8) **An obelisk** is a four-sided tapering monument with a pyramid at its top.

In example 2.7, the phrase 'an obelisk' is used to refer to a particular thing (termed the phrase's *referent*); on the other hand, the same phrase in example 2.8 has a generic interpretation (i. e., the sentence concerns the class of things called obelisks), and is not being used to refer. The terms *denotation* and *extension* are used to mean the set of

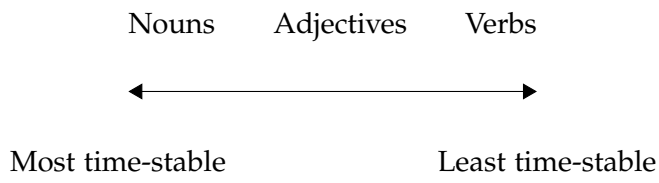


Figure 2.1: Givón's spectrum of temporal stability.

things in the world which could possibly be the referent of a given expression.

Referential or *denotational* theories of semantics place primary emphasis on the relationship of denotation; nouns can be said to be meaningful precisely because they select for classes of particular entities in the world. From this point of view, different parts of speech can be seen to have distinct denotational behaviour (Saeed, 1997, p 30):

proper names	denote individuals
common nouns	denote sets of individuals
verbs	denote actions[1]
adjectives	denote properties of individuals
adverbs	denote properties of actions

Givón (1979) proposes that the major parts of speech of language reflect a spectrum of the perceived temporal stability of the phenomena they denote. At one extreme lie 'experiences – or phenomenological clusters – which stay relatively *stable* over time, that is, those that over repeated scans appear to be roughly "the same"' (Givón, 1979, p 51). On the other extreme, we find 'experiential clusters denoting *rapid changes* in the state of the universe. These are prototypically *events* or actions' (Givón, 1979, p 52). Givón asserts that this spectrum, depicted in figure 2.1, is directly reflected in the structure of language, so that nouns denote the most temporally stable phenomena, adjectives are intermediate, and verbs denote the least temporally stable things that we use language to represent.

Further, he notes that, across languages, it seems to be the case that abstract nouns are always derived from verbs. This suggests that the noun's fundamental role in language is to denote physically located entities that are relatively stable over time, whereas verbs prototypically denote things which 'only [have] existence in time' (Givón, 1979, p 321). This distinction between the temporal and non-temporal aspects of the universe motivates the dichotomy of nouns and verbs, the two basic formal categories of language.

The kinds of things that verbs denote, then, are temporary in nature, although the situations they express may variously hold either momentarily or for a relatively long period of time. As temporality is

1 I will have more to say on the kinds of things that verbs can denote in section 2.4.

usually a central feature of verbs' denotations, these situations can be thought of as being located and ordered in time. Moreover, situations expressed by verbs frequently have internal temporal structure:

(2.9) Drenched in water, the Wicked Witch **melted**.

(2.10) Mary **began** a novel.

(2.11) Conny **finished** painting the kitchen.

In example 2.9, the verb 'melt' expresses the transition from one state into another; in example 2.10, a process starts (reading, unexpressed by the verb 'begin' but implicated by the direct object 'novel'); and in example 2.11, an activity (painting) comes to an end. It can be seen that a major part of verb meaning is to be found in their temporal or *aspectual structure* – 'how events unfold over time' (Croft, 2012, p 4). This facet of verb meaning will be investigated in further detail in section 2.4.

The second strand of verb meaning concerns the tendency of verbs to relate to other words in the same sentence (*syntagmatic relations*), alluded to in section 2.1. We now turn to this side of verb semantics.

2.3 ARGUMENT STRUCTURE

Argument structure, also known as *diathesis*, concerns the way a verb relates to the entities that participate in the situation expressed by the verb. This issue is partly related to syntactic matters, such as the grammatical relations (e. g., subject, object) that every verb manifests, and partly related to verb meaning. This section will sketch the view that a verb is a kind of predicate (section 2.3.2) before looking at how verbs impose demands on what kinds of entities must or may appear with them (section 2.3.3). The roles these entities play in situations can be characterised (section 2.3.4), which leads to a view of verbs as belonging to families expressing related meanings (section 2.3.5). At this point we will take a brief look at decompositional approaches to verb semantics (section 2.3.6) and lexical ambiguity (section 2.3.7). The section concludes by looking at the ways that verbs can pick and choose which arguments they co-occur with (section 2.3.8).

We begin by examining how sentences are assembled from their constituent parts.

2.3.1 *Compositionality*

Syntax in language allows the formation of complex expressions from simpler expressions, as when building phrases from single (or *simplex*) words, or when building sentences from phrases. The German linguist Wilhelm von Humboldt (1767–1835) may have been the first

to note that this ability to fashion larger ensembles out of collections of single words was creative, rule-based, and – most importantly for our present purpose – productive and recursive (Losonsky, 1999). This allows for the combinatoric possibility that an enumerable set of lexical items can give rise to an unlimited set of sentences; as Humboldt put it, language 'makes infinite use of finite means'. The formal and mechanical description of how words are syntactically put together into larger groups and how this recursion works was provided by Chomsky (1957), revolutionising the modern understanding of syntax.

The significance of the combinatoric nature of linguistic structure for the meaning of language was recognised earlier, however: If an infinity of complex expressions can be constructed from a finite number of lexical items, then it follows that the meanings of those complex expressions must be predictable from the meanings of their component parts, and the way in which those parts are assembled (i. e., their syntactic structure). This *Principle of Compositionality* is thus a central axiom of nearly every semantic theory; it is generally attributed to the German logician, mathematician and philosopher Gottlob Frege (1845–1925), and hence is sometimes known as Frege's Principle.

Briefly, compositionality is the idea that the meaning of a multiword linguistic expression is determined by exactly two factors: its internal syntactic structure, and the meanings of its constituents. The meanings of the constituents are, at the end of the day, explained by the lexical semantics of the words making up the expression. In the case of a verbal expression, the syntactic structure of the expression can be identified with the verb's argument structure.

2.3.2 *Predicates*

In logical semantics, *predication* is the primary mechanism whereby compositionality is achieved; it is the means by which the meanings of parts are linked together to express the meanings of the whole. This linking can happen in different ways on different levels, as can be illustrated using the following simple example:

(2.12) Monica ate an apple.

Here, the word 'Monica' is a proper name, and refers to a specific individual, a particular person named Monica. 'An apple' is a noun phrase, composed of the indefinite determiner 'an' (existentially quantifying or introducing a single entity) and the common noun 'apple'. Semantically, we can surmise that the referent of this noun phrase is an entity that is an apple. Thus, the contribution of the word 'apple' to the meaning of this phrase is that it restricts the phrase's referent to be a member of the set of things that are apples. This task is accomplished by predication, namely the application of a concept

(or *predicate*) to a particular entity (or *argument*), or possibly to many entities. Nouns, verbs, adjectives, and adverbs can all be used to predicate. In this example, letting the referent of 'an apple' be signified by a, we are told that $\textsc{apple}(a)$.

The word 'ate' is a verb, inflected to agree with the subject 'Monica' in number, as well as for the past tense. This word refers to a particular event; the past tense tells us that this event took place at some time before the moment of utterance of our sentence. The event in question involved two entities ('Monica', who performed the eating; and the entity a, which was consumed). Leaving aside tense for the moment, the word 'ate' resembles the noun in that its semantic contribution is to restrict its referent to be a kind of eating event (a member of the set of eating events). Unlike the *one-place* (or *unary*) *predicate* of 'apple', however, the verb in this sentence links both a subject and a direct object, making it a *two-place* (or *binary*) *predicate* taking two separate entities as arguments: After all, 'eating' is done by someone, and involves eating something. By representing the person Monica with the constant term m, we can write $\textsc{eat}(m, a)$.

Predicates and arguments are both modelled in semantic theories, but they differ in kind. Arguments represent entities, individuals that are independent, stand alone, and exist in the world. Predicates represent concepts that stand in relation to entities; for instance, adjectives like 'red' may denote properties, nouns like 'apple' may denote natural kinds, prepositions like 'ahead of' or comparatives like 'taller than' may denote relations, and verbs like 'eat' can denote events.

The number of arguments taken by a predicate is termed its *arity*, *adicity* or sometimes *valency* (Tesnière, 1959). *Intransitive* verbs like 'sneeze' introduce unary predicates, whose sole argument is the subject of the verb, while *transitive verbs* like 'kick', requiring a direct object, introduce binary predicates. Ditransitive verbs (e. g., 'give', 'send', 'put', etc.) take two objects and introduce *ternary predicates*. In English and German, there are also examples of *avalent verbs*, whose predicates take no arguments (common examples are descriptions of weather such as 'It's raining', where 'it' is termed an *expletive* or *null subject* and is not taken to refer to any particular entity). In *pro-drop languages*, the semantically empty subject is not required (cf. Italian *piove* 'it's raining', Spanish *llueve* 'it's raining').

Furthermore, predicates can also be classified. Arguments to nominal (noun) predicates like the apple a in example 2.12 are not introduced or created by a separate predicate in the way that the two arguments to 'ate' are; rather, a is just the referent of the noun. Because of this, such arguments are termed *referential arguments*. Langacker (1987, p 68) argues that language consists of two kinds of predication: nominal and relational. Nominal predication is expressed by nouns, and relational predications, either temporal or atemporal (e. g., spatial) in kind, are expressed by verbs or adjectives and adverbs. The

situation referred to by a verb does not exist independently in the way that an entity referred to by a noun does. Rather, the denotation of the verb is said to be *instantiated* by its arguments; its existence is a function of the temporary conditions brought about by the verb's participants. Exceptions to this rule are avalent verbs, which apply immediately to the situation referred to.[2]

2.3.3 *Subcategorisation*

We have seen how verbs may be labelled as intransitive (like 'sneeze'), transitive (like 'kick'), ditransitive (like 'give'), etc. This description reflects the traditional wisdom that, syntactically, the number and kinds of arguments a verb requires are understood to be a *lexical property*. That is, each verb may have its own peculiar requirements for the arguments that must appear with it for it to be in a grammatical utterance, and the number and types of arguments required by a verb is thought to be part of a speaker's knowledge of the vocabulary of her language. We must now further refine this picture of the syntactic needs of verbs.

Arguments to verbs do not have to be NPs; many verbs have prepositional complements, where the preposition is lexically fixed (e. g., 'depend on', 'refer to', 'differ from', etc.)[3]. Other verbs, such as verbs of motion (e. g., 'go', 'put') can be combined with a variety of prepositional phrases (PPs) indicating locations, paths, goals, etc. Still other verbs take verbal or sentential complements, such as 'that' complements ('know', 'believe', 'assume', 'say'), infinitives ('try', 'manage', 'begin to'), or gerunds ('start', 'stop', 'keep –ing').

Thus, a verb like 'put' can require, besides its subject NP, both a direct object NP, specifying the thing placed, and an *oblique PP* or suitable adverbial, specifying the location in which the thing was placed. Uses of the verb which do not meet these requirements are ungrammatical:

(2.13) (a) [$_{NP}$ Al] put [$_{NP}$ the book] [$_{PP}$ on the shelf].

(b) * [$_{NP}$ Al] put [$_{NP}$ the book].

(c) * [$_{NP}$ Al] put [$_{PP}$ on the shelf].

(d) * [$_{NP}$ Al] put.

In German, verbs may also specify the grammatical cases of their arguments; typical transitive objects are marked in the accusative case, as in example 2.14, but some verbs may require the dative (example 2.15) or even genitive case (example 2.16):

2 As we shall see shortly in section 2.4.5, event semantics proposes that verbs also have a referential argument, which refers to a particular event or state of affairs; this event argument is introduced by a finite verb's inflection for tense, aspect, and mood.

3 Examples in this paragraph are taken from Löbner (2013, p 112).

(2.14) Edi schlug Daniel.
Edi hit Daniel.ACC.
Edi hit Daniel.

(2.15) Darf ich Ihnen helfen?
May I you.DAT help?
May I help you?

(2.16) Das bedarf keiner Erklärung.
That requires no.GEN explanation.
That requires no explanation.

Verbs may also be flexible in the number or kinds of arguments they require; this flexibility is similarly an idiosyncratic property of the verb. Fillmore (2003, pp 126ff) contrasts:

(2.17) (a) John broke the stick (with a rock).
(b) A rock broke the stick.
(c) The stick broke.

with:

(2.18) (a) John hit the tree (with a rock).
(b) A rock hit the tree.
(c) * The tree hit.

While the verb 'hit' can only be transitive, we observe that 'break' exhibits considerable flexibility; it can be intransitive or transitive. Each of these instantiation patterns is called a *subcategorisation frame* (SCF); we can conclude that 'break' permits two separate SCFs, while 'hit' permits only one. The two frames admitted by 'break' also represent an instance of a *diathesis alternation*, a variation in the expressions of arguments to a verb, sometimes accompanied by a change in meaning (Levin, 1993, p 2). The particular alternation at work here is called the *causative alternation*, and permits a verb expressing a change of state to be used transitively as a *causative verb*, as in example 2.19a; or intransitively as an *inchoative verb* (a verb coding for a change of state into a new state of being), *anticausative verb*, or *unaccusative verb*, as in example 2.19b:

(2.19) (a) I opened the door.
(b) The door opened.

In syntax, a distinction is usually drawn between arguments, which are required by one of a verb's SCFs, and *adjuncts*, which are optional complements to the verb and are not assumed to be involved in the verb's predication. Adjuncts are frequently adverbial specifications of

time, place, or manner. For example, while location may be a required argument of a verb like 'put', it is an optional adjunct to a verb like 'read':

(2.20) (a) Tobi read the book in the bathroom.
(b) Tobi read the book.

(2.21) (a) Tobi put the book in the bathroom.
(b) * Tobi put the book.

'Eat', 'dance' and other verbs may have optional arguments:

(2.22) They ate. They danced.

(2.23) They ate fish. They danced a jig.

Chapter 5 is focused on empirically learning the subcategorisation frames for German verbs from a large amount of text. Prior research on subcategorisation acquisition is reviewed in section 4.1.

2.3.4 *Thematic roles*

Thematic roles allow a semantic characterisation of the relations between predicates and their arguments and adjuncts. Verbal arguments participate in the action expressed by the verb, and the thematic role of a particular argument indicates in what way the referred entity is involved. While there exist some broad patterns in how particular thematic roles are realised syntactically (e. g., the subject of the verb is often the initiator of the action), thanks to diathesis alternations and the passive voice, grammatical roles do not always align with thematic roles. Löbner (2013, p 122) gives this example:

(2.24) (a) [$_{\text{T}}$ The door] opens.
(b) [$_{\text{I}}$ This key] opens [$_{\text{T}}$ the door].
(c) [$_{\text{A}}$ The child] opened [$_{\text{T}}$ the door].
(d) [$_{\text{A}}$ The child] opened [$_{\text{T}}$ the door] with [$_{\text{I}}$ a key].
(e) * [$_{\text{I}}$ This key] opens [$_{\text{T}}$ the door] by [$_{\text{A}}$ the child].

These sentences show the verb 'open' combining with three participants: an animate agent (A) who performs the opening, an inanimate object or theme (T) that is opened, and an instrument (I) that enables or facilitates the opening. We see that the grammatical subject often codes for the agent, as in examples 2.24c–2.24d; however, when the agent is not specified, the subject can be the instrument (example 2.24b) or even the theme (example 2.24a). When it is not the subject, the theme tends to appear as the direct object of the verb (examples 2.24b–2.24d). The instrument is the subject in example 2.24b, but appears as a prepositional object in example 2.24d.

Agent	**Ozel** drove the truck.
Patient	Svetlana cleaned **the kitchen**.
Theme	Christian threw **the Frisbee** at Conny.
Instrument	Jean cut up the frame with **a jigsaw**.
Experiencer	**Natalya** heard the doorbell.
Benefactive	They cooked **me** dinner.
Source	The cat jumped off **the bed**.
Goal	Anne has moved to **Hannover**.
Location	Mary works at **the Ministry of Justice**.

Table 2.1: Examples of some commonly used thematic roles.

To better distinguish the semantic performance of entities from their grammatical function, verbal arguments are assigned thematic roles. Different semanticists employ different sets of thematic roles, and such inventories may vary in size from 18 to 25 roles (Frawley, 1992, p 201). Some commonly used roles are:

AGENT An active instigator of the action, usually human and hence animate, volitional, intentional, and causally responsible.

PATIENT A passive entity that is changed as a result of the action.

THEME An entity that moved but remains otherwise unchanged by the action.

INSTRUMENT An entity that is used to carry out the action.

EXPERIENCER An entity whose internal state is affected by the action, usually animate and sentient.

BENEFACTIVE An entity for whom the action is performed.

SOURCE The place from which an entity is moved.

GOAL The place to which an entity is moved.

LOCATION The static location where the action takes place.

Examples of these thematic roles are given in table 2.1.

As syntactic concomitants to a verb may be classified as being mandatory arguments or optional adjuncts, so thematic roles are divided into *participant roles*, which may be lexically required, and *non-participant roles*, which are always optional. The roles introduced in this section are all participant roles save for Instrument and Location; as with the argument–adjunct distinction drawn above, the category boundary is fuzzy, and some verbs may have obligatory Location roles, such as *sich befinden* 'to be located, to find oneself in a particular place'.

Chomsky (1981) introduces the 'θ-Criterion', which prescribes that each argument bear one and only one thematic role (furthermore, each thematic role can be assigned to at most one argument). By contrast, Jackendoff (1987) proposes that a given argument can have up to two thematic roles, which are assigned on two independent levels. The first level, the 'thematic tier', encodes features relating to location and motion; while the second level, the 'action tier', deals with Agent-Patient relations. In this way, the subject of the sentence 'I ran to the store' can be simultaneously an Agent (because I initiated this action and performed it volitionally) as well as a Theme (because I am in motion); similarly, the subject of the sentence 'I listened to two symphonies' can be simultaneously an Agent and an Experiencer (because my role here is described as perceiving sense data). This proposal resembles one put forth earlier by Culicover and Wilkins (1986), who distinguish 'extensional roles' related to physical location (Theme, Source, Goal, etc.) from 'intensional roles' related to an argument's status as a participant in an action (e. g., Agent, Patient, Benefactive, Instrument).

One responsibility assigned to thematic roles is variously called argument selection, argument realisation, or *argument linking* (Levin, 1993; Levin and Rappaport Hovav, 2005): Thematic roles are assumed either to fully explain, or to play a major role in how different NPs come to appear as grammatical subjects, objects, or obliques of the verb. Various linguists have proposed that each thematic role has an intrinsic affinity for the subject position, and that roles can be ordered by the strength of this affinity: If an Agent is mentioned as a participant in some action, of course it must manifest as the grammatical subject; only if an event is presented without naming an Agent may some other thematic role occupy the subject relation (this explains why example 2.24e is grammatically malformed). For example, Bresnan (2001) proposes the following thematic role hierarchy:

> Agent > Recipient > Experiencer/Goal > Instrument > Patient/Theme > Location

By contrast, Dowty (1991) proposes *thematic proto-roles*, which are defined by a set of lexical entailments. Proto-agents tend to (p 572):

- be volitionally involved in the event;
- be sentient and perceptive;
- cause a change of state in another participant; and
- be in motion relative to another participant.

Proto-patients, in contrast, tend to:

- undergo a change of state;

- be causally affected by another participant;
- be stationary relative to another participant; and
- be 'incremental themes', a topic I will come back to in section 2.4.7.

The thinking here is that the participant in an event with the most proto-agent properties will manifest as subject, and the participant with the most proto-patient qualities will assume the role of the direct object. Participants which have similar counts of entailments are predicted to be able to manifest as either subject or object.

The applied NLP task of automatically determining which thematic role a syntactic argument to a verb has is called *semantic role labelling* (SRL), and will be described in more detail in section 3.4, along with efforts to make concrete the idea of thematic roles in the form of annotated corpora. In section 8.2 I present an experiment using automatic SRL as a task to determine the quality or usefulness of a supervised classifier.

A recent overview of work on thematic roles is given by Davis (2011).

2.3.5 *Verb classes*

As discussed in section 2.3.3, verbs may require different combinations of arguments. These arguments, which we have so far characterised syntactically, can equally well be described in terms of their thematic roles. Thus, a sentence like 'I sold my car to him' can be abstracted to the set of participant thematic roles involved in the predication:

(2.25) sell: <**Agent**, Theme, Recipient>

This example uses the role of Recipient, which is sometimes distinguished from a Goal specifically for events involving changes of possession (Andrews, 1985, p 70). The Agent role is set in bold face here to indicate that it is the subject of the verb. A representation of a group of thematic roles like this is called a *thematic role grid* or sometimes *theta grid*.

Specifying the thematic role grids for multiple verbs reveals that verbs form classes that share grids, and that are also related in meaning. Other verbs sharing the theta grid shown in example 2.25 include 'give', 'lend', 'supply', 'pay', 'donate', 'contribute', 'rent', etc.; all these verbs denote a transfer of possession and foreground the role of the agent. There is a similar class of verbs denoting transfers of possession, where the role of recipient is emphasised:

(2.26) buy: <**Recipient**, Theme, Source>

Verbs patterning like 'buy' include 'receive', 'accept', 'borrow', 'purchase', 'rent', 'hire', etc.

Levin (1993) assigns verbs to classes by testing whether they enter into diathesis alternations such as the causative alternation introduced in section 2.3.3, example 2.19. She assembles a list of around 80 alternations; examples include the *middle* and *conative* constructions, which look like this[4]:

(2.27) (a) Margaret cut the bread. (Transitive construction)

(b) The bread cuts easily. (Middle)

(c) Margaret cut at the bread. (Conative)

We observe that 'cut' enters into the transitive construction and these two alternations. Other verbs, however, do not:

(2.28) (a) Janet broke the vase.

(b) Crystal vases break easily.

(c) * Janet broke at the bread.

(2.29) (a) Terry touched the cat.

(b) * Cats touch easily.

(c) * Terry touched at the cat.

Equipped with these alternations, Levin checks over 3,000 English verbs, recording which alternations they permit, and grouping together those verbs that pattern similarly. The result is a list of 49 verb classes, some of which include more specific sub-classes (giving 192 classes in total). The verbs in any given class carry related meanings and express the same thematic role grids. For example, the set of verbs which enter the same alternations as 'break' (the Break class) includes 'chip', 'crack', 'crash', 'crush', 'fracture', 'rip', 'shatter', 'smash', 'snap', 'splinter', and 'tear'. Levin argues that the meaning of a verb determines, or at least strongly influences, its argument structure.

Levin's verb classes have been developed into a popular machine-readable dictionary called VerbNet; this resource is introduced and discussed in greater detail in section 3.5.

2.3.6 *Semantic components*

Decompositionality is a name given to a series of approaches to semantics, whose common idea is that the word is not the smallest unit of meaning. Rather, words are postulated to have complex meanings that are built up of sub-word *semantic components*. Semantic components have been used to try to explain regular lexical relations such as hyponymy and antonymy. Given an analysis like the following:

4 Examples from Levin (1993, p 6).

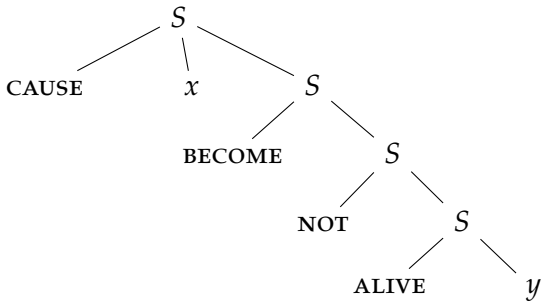


Figure 2.2: A decompositional representation of the sentence '*x* kills *y*' proposed by McCawley (1968).

(2.30)

$$\begin{aligned}
\textsc{man}(x) &\rightarrow \textsc{adult}(x) \wedge \textsc{male}(x) \wedge \textsc{human}(x) \\
\textsc{husband}(x) &\rightarrow \textsc{adult}(x) \wedge \textsc{male}(x) \wedge \textsc{human}(x) \wedge \textsc{married}(x) \\
\textsc{bachelor}(x) &\rightarrow \textsc{adult}(x) \wedge \textsc{male}(x) \wedge \textsc{human}(x) \wedge \neg\, \textsc{married}(x)
\end{aligned}$$

it seems simple enough to deduce that 'bachelor' and 'husband' are both hyponyms of 'man', and also to conclude that 'bachelor' and 'husband' are in some way opposed to each other or incompatible in their meanings.

The 1960s program of generative semantics was the first semantic framework that represented the meaning of verbs using meaning components. In particular, McCawley (1968) stands out, who put forward CAUSE as an abstract verb in the deep semantic structure of a sentence, indicating the causal relationship between an agent and a caused event; his famous analysis defines the verb 'kill' periphrastically as 'to cause to become not alive', positing that the sentence '*x* kills *y*' would have a *deep structure* (in transformational grammar, the underlying syntactic-semantic stuff from which real sentences are derived and realised) that would resemble something like what is depicted in figure 2.2. In section 2.4.6 below I shall return to decompositional approaches to verb semantics in this vein by Lakoff (1972) and Dowty (1979).

Levin and Rappaport Hovav's work on verb classes, introduced in the previous section, has also gone in a decompositional direction, leading to a program called Lexical Conceptual Structure (Levin, 2011). In their view, a verb's ability to participate in a particular diathesis alternation is contingent on the verb having a particular internal semantic structure. In other words, the common characteristic of verbs in a verb class is that they all share one or more meaning components (such as MOTION, CONTACT, CAUSE, etc.).

We shall see in section 2.3.8 that semantic components like those shown in example 2.30 are also used by Katz and Fodor (1963) to give an account of selectional restrictions, the tendency of a verb to select that certain kinds of words be its arguments.

For more background on semantic components, see Engelberg (2011b,a).

2.3.7 *Lexical ambiguity*

The discussion up to this point has bypassed a perennial complication of lexical semantics, which is that a given word form may be used in different contexts to carry different meanings. To some extent, such differences in meaning may be due to a certain flexibility of the word's denotation; in other cases, however, it can be argued that a particular word form should be thought of as having multiple, independent *word senses*. Cruse (1986, p 51) illustrates this distinction using the following examples:

(2.31) Sue is visiting her cousin.

(2.32) We finally reached the bank.

In example 2.31, the word 'cousin' can denote a person who is either a man or a woman, but the sentence can be successfully interpreted without the hearer having to choose one or the other gender; this happens because the word 'cousin' carries a *vague* or *underspecified* meaning that is compatible with both male and female gender.

By contrast, in example 2.32, 'bank' can signify either the edge of a river or a financial institution, but the word has no meaning that covers both of these interpretations. As a result of the *ambiguity* of the word, the hearer must choose one of these senses when interpreting the sentence. If the sense chosen by the hearer is not the one intended by the speaker, the sentence may fail at communicating the speaker's intent.[5] This phenomenon of ambiguity in word meaning is termed *polysemy*, and is reflected in the familiar structure of dictionaries, where a single lexical entry can list multiple senses. In the exact use of the term, polysemy is usually distinguished from *homonymy*, such that the major groupings listed in a dictionary are homonyms, words which are spelled alike but which have different etymologies (e. g., the 'bark' of a tree and a dog's 'bark', Krovetz, 1997); and the minor groupings underneath a single homonym are termed word senses or polysemes.

Polysemy is common in natural languages, and particularly so for verbs. Fellbaum (1990, p 40) reports that '. . . the Collins English Dictionary lists 43,636 different nouns and 14,190 different verbs. Verbs are more polysemous than nouns: The nouns in Collins have on the average 1.74 senses, whereas verbs average 2.11 senses.' She surmises that this greater degree of polysemy indicates that verbs are semantically more flexible than nouns.

5 Note that the word 'cousin' has only a single word sense and is not ambiguous: It is *monosemous*. This single sense is also underspecified for the degree of relation; hence, we speak more precisely of 'first cousins', 'second cousins once removed', etc.

Word sense disambiguation (WSD) is the applied NLP task of automatically determining which sense of a polysemous word is intended in a particular context.

Kennedy (2011) presents a recent survey of work on lexical ambiguity.

2.3.8 *Selectional preferences*

Selectional preferences refers to the intuition that not every predicate can be equally well satisfied by every particular argument. As an example, we can imagine that the verb 'eat' is more likely to appear with certain objects than others:

I eat	breakfast
	sandwich
	apple
	⋮
	newspaper

An early account of this is given by Katz and Fodor (1963), who set out to give a mechanical description of a semantic theory by splitting the process of interpreting an utterance into two parts: retrieving lexical descriptions from a kind of dictionary, and then combining the various concepts according to compositionality rules. A problem arises from polysemy, whereby words with more than one sense multiply the number of sentential interpretations. This combinatoric growth can be somewhat stemmed by allowing linguistic contexts to select semantically admissible word senses. To formalise this, Katz and Fodor associate lexical entries in the dictionary with a series of decompositional *semantic markers* (e. g., Animal, Human, Male, Female, Colour, Weight, etc.); the non-systematic part of a word meaning that cannot be expressed using markers is indicated using the sense's *distinguisher*. Lexical entries for predicates can specify markers which must be held by arguments in order to be processed by the compositionality rules (e. g., Subject: Human, or Object: Physical Object); this kind of constraint on predication is termed *selectional restrictions*.

Wilks (1975) imagines a more flexible scenario, where predicates do not insist on selectional restrictions being satisfied, but rather merely express a preference that they do so. All other things being equal, this mechanism should still arrive at the same choice of word senses for an ambiguous sentence as Katz and Fodor would; however, the extra permissiveness allows language to be interpreted even when it violates selectional restrictions. Wilks (1978) later extends this model to identify metaphorical constructions because these frequently violate selectional restrictions (e. g., 'My car drinks gasoline').

An even more modern view is taken by Resnik (1993), who models selectional preferences probabilistically: Here, arguments are associated with various conceptual classes, and predicates may be modelled as having a stronger or weaker association with particular classes of arguments.

As Quirk et al. (1985, pp 771f) note, selectional restrictions may apply to categories such as number (example 2.33), concreteness (example 2.34), animacy (example 2.35) and humanness (example 2.36):

(2.33) (a) The men scattered.
(b) * The man scattered.

(2.34) (a) The glass contains water.
(b) * The glass contains kindness.

(2.35) (a) A pedestrian saw me.
(b) * A lampshade saw me.

(2.36) (a) We got married.
(b) * The snakes got married.

Chapter 6 is concerned with modelling the selectional preferences of German verbs using corpus data. A survey of prior research on selectional preference methods is presented in section 4.3.

2.4 ASPECTUAL STRUCTURE

Aspectual structure concerns the shape of an event in time. Verbs refer to different kinds of situations, which can be states or dynamic processes or events. The lexical meaning of a verb is to a large degree concerned with how the action signified by the verb unfolds in time, how long it is, whether it produces changes in any of the participants in the event, etc. Section 2.4.1 starts out by distinguishing grammatical aspect, a syntactic category, from lexical aspect, the semantic category that I am chiefly concerned with. Section 2.4.2 looks closer at lexical aspect, suggesting that there are a small number of different aspectual types of verbs, and section 2.4.3 introduces relevant aspectual categories, capturing temporal features of verbs. These ideas can be put together to create classifications of lexical aspect; section 2.4.4 surveys some of these typologies. Section 2.4.5 briefly introduces event semantics, and section 2.4.6 discusses the notion of change as it relates to lexical aspect. The final two sections are dedicated to refining this view of lexical aspect, by discussing the role of the verb's Patient (section 2.4.7) and sketching how the lexical class of a verb can be modified by its linguistic context (section 2.4.8).

2.4.1 *Grammatical aspect*

The term 'aspect' in the context of linguistics is often used to refer to a grammatical category that allows an event to be presented with either a perfective or an imperfective viewpoint. I shall follow Klein (1994) here by naming this syntactic category more precisely as 'grammatical aspect', briefly sketched here, to better distinguish it from 'lexical aspect', a semantic category addressed in the next section, where our real focus lies.

Grammatical aspect is concerned with how an event, which occurs over a particular span of time, is described to the hearer, prototypically using either the *perfective* or *imperfective* aspect to do so:

(2.37) Axel **was reading** a book when he had an idea. (imperfective, progressive)

(2.38) Axel **went** straight home again. (perfective)

(2.39) Axel **is reading** *Ulysses*. (imperfective, progressive)

(2.40) Axel **was running a mile a week** back then. (imperfective, habitual)

The perfective aspect, illustrated here with example 2.38, may be used to present

> the totality of the situation referred to ... without reference to its internal temporal constituency. The whole of the situation is presented as a single unanalysable whole, with beginning, middle, and end rolled into one; no attempt is made to divide this situation up into the various individual phases that make up the action of entry. (Comrie, 1976, p 3)

By contrast, the imperfective aspect focuses on the innards of an event, allowing the event to be used as the backdrop to temporally locate some other episode, as in example 2.37.

Comrie (1976) proposes a classification of the sub-categories of grammatical aspect, which is sketched in figure 2.3. Imperfective verb phrases can be further subcategorised as having either *habitual* or *continuous* aspect. Habituals indicate a tendency to perform a particular action repeatedly over a period of time, as in example 2.40; habituals can also refer to states: 'The Temple of Diana used to stand at Ephesus' (Comrie, 1976, p 27).[6]

6 Habituals in the past tense have an implicature that the situation described no longer holds in the present (cf. Mitch Hedberg's 'I used to do drugs. I still do, but I used to, too.').

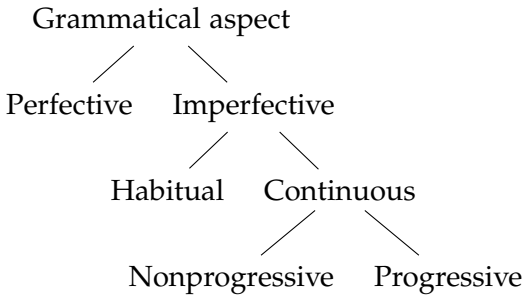


Figure 2.3: A classification of grammatical aspect adapted from Comrie (1976, p 25).

Continuous verb phrases, by contrast, refer to a single process over a given time period. A sub-class of these, *progressive* verb phrases describe the performance of a process, as in examples 2.38 and 2.39. Because of its focus on the interior of a time span, the progressive does not convey any detail on the beginning and end points of the event, which means that it implies that the action is not completed. Example 2.38 shows this effect, too. From the sentence as written, one cannot logically conclude that Axel finished his book; rather, his reading is explicitly interrupted by a moment of inspiration. Non-progressive aspect, a feature of many languages (although not English or German), applies to verb instances marked for the imperfect that refer to states rather than actions. Sometimes a given expression is ambiguous for aspect, as with example 2.39, which can be either ongoing (Axel is currently turning pages) or continuous and iterated (Axel has started the book and intends to finish it sometime).

Note that features such as iterated or habitual aspect are also treated from a semantic point of view; the category of aspect can be seen as a single edifice straddling the syntax-semantics boundary, with grammatical aspect referring to those qualities that are grammaticalised in the language's syntax. In *grammaticalisation*, a semantic category becomes cemented in the grammar, such that the language compels the speaker to make a distinction when constructing a sentence.

The dichotomy of 'perfective' and 'imperfective' was originally coined to describe Slavic languages, where this feature is morphologically marked on the verb in all tenses. Latin, and hence Romance languages more generally, distinguishes between the perfective and imperfective only in the past tense. English grammar distinguishes between verbs with the progressive aspect and verbs without it, which in turn signal the imperfective and perfective aspects respectively; English also grammaticalises the habitual aspect in the past tense: 'He used to work here.'

German does not grammatically distinguish perfective and imperfective, even though it has two past tense forms and thus would be capable of doing so. This is because many dialects of German no

longer make use of the simple past, and have replaced it entirely with the present perfect. There are constructions that can indicate imperfective aspect, such as *Er las das Buch* 'He read the book' modified to *Er las im Buch* lit. 'He read in the book', with the meaning that the person in question did not finish reading the whole book. However, this device is only compatible with a small number of verbs. A further imperfective construction in German is the *Rheinische Verlaufsform* ('Rhenish progressive') or *am-Progressiv* ('*at* progressive'), e. g., *Ich bin am Arbeiten* (lit. 'I am at the work'). As we shall see in section 7.2, this construction turns out to be very rare.

2.4.2 *Lexical aspect*

Semantically, verbs are often initially sketched as representing the main action in any given sentence (cf. the German grade school term *Tunwort*, 'doing word').

In fact, many verbs do not represent actions at all. In this example:

(2.41) Sara knows how to ski.

neither is Sara actively doing something, nor is there some ongoing process involving Sara or skiing. Rather, the verb 'know' in this context is better described as ascribing some property to Sara (that she possesses the ability to ski); the present tense of the sentence tells us that this property holds (or *obtains*) at the present moment (the moment of utterance), but sentence does not indicate when the property came about, or otherwise betray any details about the temporal development of Sara or her skiing prowess. 'Know' is an example of a *stative verb*, and merely expresses a static, unchanging attribute of its subject.[7] By contrast, *dynamic verbs* are those that are not stative, referring to processes, events, activities, etc.

Dowty (1979, pp 55f) lists the following tests for stative verbs:

1. They are incompatible with the progressive:

   (2.42) I am searching for a new apartment.

   (2.43) * I am knowing English very well.

2. They cannot be complements to the verbs 'force' and 'persuade':

   (2.44) She forced me to move out.

   (2.45) ? She forced me to be tall.

3. They cannot occur in the imperative:

   (2.46) Go home!

7 While the straightforward identification of actions with verbs fails, the intuition is not entirely misplaced: When actions are expressed in language, they are almost always represented using verbs.

(2.47) ? Know the answer!

4. They do not combine with adverbs 'deliberately', 'carefully':

   (2.48) I deliberately read the instruction manual.

   (2.49) ? I deliberately knew the answer.

5. They do not enter into the pseudo-cleft:

   (2.50) What I did was to cross the street.

   (2.51) ? What I did was to know the answer.

6. The present tense usually evokes a habitual interpretation for dynamic events, but not for states:

   (2.52) I live in the city. (state obtaining at the moment of utterance)

   (2.53) I play tennis at the club. (habitual)

This leads us to the concept of *lexical aspect*, sometimes called *Aktionsart*. This is a semantic category that is projected by the verb; as we shall see, the aspectual type of a verb phrase may also be affected by the verb's arguments and modifiers. A recent survey of work on lexical aspect is to be found in Filip (2012).

2.4.3 *Aspectual categories*

An aspectual feature of some dynamic verbs is the notion of *telicity*; telic verbs are also sometimes called 'resultatives'. It was already noted by Aristotle in his *Metaphysics* that some verb phrases refer to processes that have an obligatory end: If I am making a chair, then, barring major misfortune, at some point in the future, I will have produced a chair. After this point in time, it is no longer felicitous to describe maintenance or modification as 'making the chair', because that particular act of making is fully accomplished at this point. On the other hand, if tragedy does strike and I am prevented from finishing my work, I also cannot describe my incomplete efforts as a 'chair-making' event, the essential and sufficient criteria being that the action signified by the verb was allowed to run its course to its natural end, and that an artefact worthy of the name 'chair' was produced as a result.

Several tests for telicity exist, including that the progressive does not entail the simple past (Kenny, 1963):

(2.54) 'I am now looking at the sunset' entails 'I have looked at the sunset'. (atelic)

(2.55) 'I am now building a house' does not entail 'I have built a house'. (telic)

Related is the notion of interruption:

(2.56) 'I stopped looking at the sunset' entails 'I looked at the sunset'.

(2.57) 'I stopped building the house' does not entail 'I built the house'.

In discussing telic phenomena, I will at times use the term *boundedness* to refer specifically to the intuition that some processes can continue to run indefinitely, while other events have a natural conclusion, after which the action is complete. The other central quality of telicity is a sense of change, as we shall explore in section 2.4.6.

The difference between unbounded and bounded events can be observed in the combination with durative adverbials such as 'for ten minutes' or 'all night long', as well as time-span adverbials such as 'in ten minutes':

(2.58) Conny looked out the window for ten minutes.

(2.59) ? Conny looked out the window in ten minutes.

(2.60) ? Conny baked a dozen muffins for ten minutes.

(2.61) Conny baked a dozen muffins in ten minutes.

Unbounded events, like looking out of a window, can continue forever. The combination of a stative or unbounded event with a durative adverbial is allowed[8]; however, time-span adverbials are infelicitous. The pattern is reversed with bounded events. Baking some muffins is a process that must necessarily stop after some period of time, which is compatible with the time-span adverbial. The durative adverbial applied in example 2.60 is difficult to interpret, but could be taken to imply that the muffins might not have been baked to completion.

Jackendoff (1983) looks at boundedness as a kind of semantic component, so that bounded events are identified with countable nouns and bounded paths; unbounded events are much like bare plural NPs. He gives this example (p 246):

(2.62) Oil was leaking $\left\{ \begin{array}{c} \text{onto} \\ \text{all over} \end{array} \right\}$ the floor.

(2.63) Some oil was leaking $\left\{ \begin{array}{c} \text{onto} \\ \text{?? all over} \end{array} \right\}$ the floor.

(2.64) People were running all over the place. (unbounded)

8 An unbounded event combined with a durative adverbial produces a bounded event by transformation, as we shall see below in section 2.4.8.

(2.65) Some people were running all over the place. (bounded)

The mass noun 'oil' in example 2.62 refers to an unbounded substance. The application of the quantifier 'some' in example 2.63 transforms this into a bounded substance, a physical object with spatial boundaries; this change makes for the semantic oddness of 'all over'. Similarly, bare plural NPs denote unbounded collections and are uncountable, acting like a substance. In example 2.64, an amorphous mass of people is spread out and covers a particular area homogeneously; with a bounded collection, by contrast, several individual people are each running aimlessly through some space. Krifka (1987) formalises this identity between mass or count nouns and unbounded or bounded events.

A second feature of the lexical aspect of dynamic verbs is the dichotomy between *durativity* and *punctuality*, advocated by Verkuyl (1972). Durative (or extended) processes last for a period of time, whereas punctual events (sneezing, recognising your mother, finding your keys, crossing the finish line) appear to occur instantaneously. When combined with durative adverbials, punctual events behave differently than extended events:

(2.66) Patti slept.

(2.67) Patti slept all night.

(2.68) Patti was sleeping.

(2.69) Patti coughed.

(2.70) Patti coughed all night.

(2.71) Patti was coughing.

In example 2.70, it cannot be the case that Patti generated one single cough after a whole nighttime of struggling to produce it; rather, the durative adverbial induces an interpretation that the punctual event was *iterated* or repeated multiple times. The same holds of the application of the progressive in example 2.71.

2.4.4 *Aspectual classifications*

Vendler (1967) provides the first classification of lexical aspect. He distinguishes four aspectual classes of verbs: states (statives; e. g. 'know'), activities (extended and atelic; e. g., 'push a cart'), accomplishments (extended and telic; e. g., 'paint a picture') and achievements (punctual and telic, e. g., 'recognise').

Dowty (1979) lists several tests for membership in Vendler's classes. For instance, accomplishments can be complements to the verb 'finish', but achievements are not compatible:

(2.72) He finished reading his book.

(2.73) ? He finished finding/noticing his book.

Accomplishments can be combined with both 'for'- and 'in'-adverbials, while achievements can pair with 'in' but not with 'for':

(2.74) He read his book for five minutes / in five days.

(2.75) He found his book in five minutes / ? for five minutes.

Finally, as first noted by Ryle (1949), there exist adverbs of intentionality or agentivity ('attentively', 'studiously', 'vigilantly', 'conscientiously', 'obediently', 'carefully') that work with activities and accomplishments but are incompatible with achievements:

(2.76) He obediently parked the car.

(2.77) ? He carefully reached Paris.

(2.78) ? He intentionally noticed the painting.

Smith (1991) introduces a fifth category to Vendler's classification: *semelfactive verbs*. These are punctual and atelic verbs such as 'cough', 'knock', 'sneeze', 'blink', etc. They are particularly amenable to the iterative reading when combined with durative adverbials (e. g., 'the light was flashing for five hours'); telic predicates are more strange in this combination (? 'I was/kept finding my keys for five hours').

Egg (2005) argues that the iterated reading of a punctual event when combined with a durative adverbial can be explained simply as a mismatch between the expected duration of the event and the time span reported by the speaker. In 'Joe sneezed for five minutes', it is the opposition between the typical duration of a sneeze (on the order of a second) and the specified five minutes that produces the sense of iteration. As a consequence of this reductive explanation, Egg's classification of aspectual types does away with the punctual–extended dichotomy, viewing this as a gradual cline and not an exact categorical distinction.

2.4.5 *Event semantics*

Much of the work described up until this point has more or less identified the boundedness of events with their telicity. A different view on aspectual semantics is afforded using ideas from event semantics.

Event semantics is traced to Davidson (1967), who postulated the existence of a separate kind of ontological thing in the universe of formal semantics, separate from entities, times and places: an event. Events were a useful formal device for dealing with tense semantics and adverbial modification of verbs. Particular instances of events

are tied to a particular time and place, so a sentence describing an event introduces it with the existential operator, as in example 2.80 (to paraphrase, there exists an event e such that e was a buttering event performed by Jones on the toast, and that e was performed with a knife):

(2.79) Jones buttered the toast with a knife.

(2.80) $\exists e.\textsc{butter}(\text{Jones}, \text{toast}, e) \wedge \textsc{instrument}(\text{knife}, e)$

(2.81) $\exists e.\textsc{butter}(e) \wedge \textsc{agent}(\text{Jones}, e) \wedge \textsc{patient}(\text{toast}, e) \wedge \textsc{instrument}(\text{knife}, e)$

Neo-Davidsonian event semantics offloads further work onto the event argument, pulling semantic roles out of the main predicate and attaching them instead directly to the event (example 2.81). It also postulates event arguments for states, as well as for adjectives, *deverbal nouns* (e. g., 'the destruction of the city') and often other kinds of predicates, too.

Bach (1981, 1986) grounds event semantics more formally by defining a model-theoretic *mereology*, a way to represent part-whole relations. Link (1983) had introduced mathematical 'lattice' structures to model the logic of physical collections and substances, as referred to by bare plurals and mass nouns. Bach extends this structure to describe events and parts of events, ordered in time. Bach divides aspectual types into three classes: states, unbounded 'processes', and bounded 'events'. He uses the umbrella term 'eventuality' to refer to a member of any of the three classes.

Maienborn (2011) provides a more detailed survey of the field of event semantics.

2.4.6 *Change of state*

A result of work on event semantics was the development of decompositional analyses of aspect affording a greater emphasis on the idea of change as central to telicity. Wright (1963) sets out a definition of a change-producing event (e. g., closing a door) as the replacement of one state (door open) by another (door closed) at a particular moment in time; he argues that any change event can be described as the transition of some particular state from false to true. Lakoff (1972) reworks this into BECOME, an operator indicating a change of state, which is predicated of (stative) sentences. Dowty (1979) proposes a decompositional calculus of aspect, whose syntax is built up from states combined with aspectual operators. The $\textsc{become}(p)$ operator encodes a change of state from $\neg p$ to p; CAUSE is used for accomplishments to demonstrate causation; and DO is for agentivity, related to intentionality or volition. Like Bach (1981), Dowty conceives of lexical

aspect as classifying verbs into being states, unbounded processes, or bounded events (which use the BECOME operator).

Some classes of telic verbs, under this view, necessarily involve a change of state in the world; they denote events with a built-in goal, where, after this target is achieved, the action is naturally complete. Another way to express this is to say that telic events with a change of state have a climactic instant in time, after which some state, the *post-state*, holds. Typical examples of change of state include the creation or destruction of something, or some other change applied to the Patient of the verb, or else motion along a path that reaches a specified goal. Nedjalkov and Jaxontov (1988) argue that telic events are precisely those whose post-states necessarily entail that some previous event occurred. In the example (p 7):

(2.82) John has broken a stick.

(2.83) The stick is broken.

the sentence example 2.82 logically entails the sentence example 2.83. In order for example 2.82 to be an accurate description of an event, it must have been the case that example 2.83 was at first a false statement, and then some event occurred, which made example 2.83 true.

Change of state verbs that are durative have a *preparatory process* that occurs and brings about the change event; modification by a durative adverbial indicates the amount of time needed for this preparatory process. When punctual change of state events are combined with a durative adverbial, the natural interpretation is that the time span is predicated of the post-state (e. g., 'I left the room for an hour' means that the change introduced – not being in the room – held for the named time interval).

Change of state verbs presuppose that their post-state does not hold before they occur:

(2.84) ? I killed him after he was already dead.

2.4.7 *Aspectual consequences of verbal arguments*

As already intimated several times in the previous sections, the aspectual class of a verb phrase can be influenced by the verb's Patient or Theme. As already pointed out in section 2.3.4, the Patient of a verb is prototypically its direct object; however, the Patient can also manifest as the subject for unaccusative verbs (Perlmutter, 1978; Levin and Rappaport Hovav, 1995).

We have seen in section 2.4.3 that indefinite plural NPs or mass nouns refer to unbounded things (cumulative substances, and quantised collections). When accomplishment predicates are combined with this kind of nominal, they become a Vendlerian activity, an unbounded event (Verkuyl, 1972; Dowty, 1979):

(2.85) John built a house. (accomplishment)

(2.86) John built houses. (activity)

The significance of the verb's Patient to its aspectual class is also related to the idea of *incremental theme* (Tenny, 1987; Dowty, 1991). This can be understood as a kind of thematic role, a property holding of an argument to a predicate, for which parts of the object signified by the argument can be mapped directly onto parts of an event predicated of that argument. For example, in:

(2.87) Matthias drank a glass of beer.

the event described took place over a certain span of time; parts of the glass of beer are drunk by Matthias during sub-spans of this total event time span. The mapping between parts of the glass of beer and parts of the time span is homomorphic (structure-preserving) in the sense that the pieces of beer together constitute the whole glass, and the pieces of time are together all contained in the whole time span. None of the sub-events where a portion of beer is consumed during a part of the event time span can be properly described as 'drinking a glass of beer', but the sum of all these infinitesimal sub-events together can be. Krifka (1992) gives a formal definition of incremental theme with a Bach-style mereological model of time spans, and uses higher order predicates to capture the differing quantificational implications of bounded and unbounded arguments. Krifka (1998) argues that whether a verb phrase is telic or atelic is directly determined by the boundedness of the verb's argument, either in the form of incremental theme (e. g., 'eat an apple'), space (e. g. 'walk from the university to the capitol'), or quality (e. g., 'bake the lobster').

Other verbal adjuncts may also influence the aspectual class of a clause. For example, the combination of a verb of motion and a prepositional phrase coding for a bounded path leading to a Goal gives an accomplishment:

(2.88) walking (unbounded)

(2.89) walking to the park (bounded, telic)

2.4.8 *Aspectual transformation and coercion*

As we have seen, verbal arguments can change the aspectual class of a verb phrase. The same is true of a series of linguistic constructions, which can be collectively called *aspectual operators*. For example, a durative 'for'-PP can turn a unbounded event into a bounded one:

(2.90) I gazed at the sunset. (unbounded)

(2.91) I gazed at the sunset for ten minutes. (extended no change)

Similarly, some verbs act as aspectual operators (e. g., 'begin', 'finish', 'complete', 'continue'):

(2.92) I stopped gazing at the sunset. (extended no change)

The task of deriving the aspectual form of a sentence from the aspectual type of its main verb combined with any aspectual consequences from the verb's arguments or modifiers is made significantly more difficult because events seem to change their aspectual shape to meet the pragmatic needs of the sentences they are described in. As Dowty (1979, p 61) laments, 'I have not been able to find a single activity verb which cannot have an accomplishment sense in at least some special context.'

Moens and Steedman (1988) address this issue by presenting a mechanical theory of aspectual types, operators, and coercion. They assume as a description of an event a prototypical structure called a 'nucleus'; this consists of a preparatory process, 'goal event' or 'culmination', and a 'consequent state' that holds after the event is culminated. In their example (p 16):

(2.93) Harry reached the top.

we see an instance of a 'culmination': a punctual event that is accompanied by a change of state, after which some particular consequent state obtains (being at the top). Even though the culmination only explicitly codes for a change of state and a post-state, the hearer can rather easily imagine a preparatory process associated with the event; it is this part of the action (i. e., climbing perhaps) that is focused when the sentence is placed in the progressive:

(2.94) Harry was reaching the top.

Moens and Steedman's classification of aspectual types distinguishes stative from dynamic verbs; dynamic verbs are then further cross-classified by punctuality/durativity and change-of-state/no-change-of-state. The resulting five-way classification lists five aspectual classes:

1. States: indefinitely extending states of affairs; there are states that are lexical in nature (e. g. 'know'), and there are also states corresponding to the habitual, the progressive, and the consequent or post-state of an event;

2. Points: punctual events with no change of state (e. g., 'hiccup', 'tap');

3. Processes: durative events with no change of state (resembles Vendler's activity class: 'walk', 'push a cart');

4. Culminated processes: durative events with a change of state (resembles Vendler's accomplishments: 'eat a sandwich');

5. Culminations: punctual change events (resembles Vendler's achievements: 'win the race').

Members of different aspectual classes may exhibit different behaviour when combined with various aspectual operators. Moens and Steedman give the example of points and culminations, combined with the present perfect tense:

(2.95) Harry has reached the top. (culmination + present perfect)

(2.96) ? John has hiccuped. (point + present perfect)

Example 2.96 is semantically odd, Moens and Steedman argue, because the perfect tense demands that its argument be a culmination. When the predicate hiccuping is used in the present perfect, the hearer is asked to imagine a scenario where the act of hiccuping produces a change in the world. In the given example, this is not easily done, which explains why the sentence resists interpretation.

According to Moens and Steedman, aspectual operators such as tense, grammatical aspect, temporal and aspectual adverbials 'transform' the aspectual type of a verb phrase:

- the progressive construction requires a process, and produces a progressive state;
- 'for'-adverbials require processes, and result in culminated processes;
- perfect tenses need culminations, and the result is that the culmination's consequent state is asserted to obtain at the utterance's reference time; and
- 'in'-adverbials take culminations, and give culminated processes by emphasising and quantifying the preparatory process associated with the change of state.

Operators combined with arguments that are not of the right type force *aspectual coercion*, where the event being described is reinterpreted in a way that alters its aspectual class.

- Culminations can be coerced to culminated processes by bringing a focus on a preparatory stage that enables the culmination to occur;
- culminated processes can be coerced to processes by removing the implication of culmination; this explains the sense of incompleteness delivered by sentences like 'Conny was baking a cake';
- similarly, processes can be coerced to culminated processes by adding an implication that the process has been done to completion;

- points can be coerced to processes via iteration, either explicitly with adverbials like 'every week' or 'often', or implicitly as we have seen with 'Patti was sneezing all night';
- points can be coerced to culminations if the hearer can identify some relevant consequence of the punctual event; and
- any type of event can be coerced to a point by 'zooming out' on the event, viewing it as an atomic and inconsequential occurrence.

Moens and Steedman show how regular aspectual ambiguity can arise because of the multiple paths possible through the aspectual class network between two nodes, as long as the pragmatic demands of the type coercion are met. They also demonstrate how the cyclical nature of the network allows a speaker to produce arbitrarily complex nested aspectual types, building up to their worked example:

(2.97) It took me two days to play the "Minute Waltz" in less than sixty seconds for more than an hour.

2.5 BUT WHAT ABOUT . . . ?

Because of its focus on lexical semantics, this dissertation does not treat every syntactic and semantic category on the verb. This section introduces two of these dimensions, tense and modality, and argues that they are not lexically conditioned in the same way that argument structure and lexical aspect are.

2.5.1 *Tense*

Tense is a grammatical category that is marked on finite verbs in most languages, although there are examples of tenseless languages such as Burmese (Comrie, 1985, pp 50f). The function of tense is to locate situations and events in time; while it is the verb that is inflected for tense, the location in time expressed by tense is applied to the whole proposition described by the sentence or utterance. This locating function can also be carried out using adverbials of time such as 'yesterday' (*deictic*: the referent of this word is fixed by context) or 'before' (*anaphoric*: constructions like these can be used to point back to referents that were introduced earlier in the discourse).

Language locates moments and intervals by specifying their position along a time line, pictured in figure 2.4; this line is a unidimensional extent that stretches out from a central reference point representing the now. The present on the time line is actually modelled as a small nonempty interval; this present interval serves to divide the time line into two other segments, representing the past and

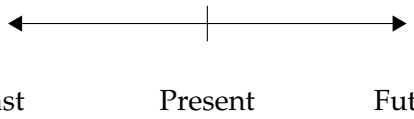


Figure 2.4: The time line in language.

future. *Absolute tense* specifies the time of an event relative to the current time (or the deictic 'origo', cf. Bühler, 1982), also known as the *moment of utterance*, or *tense locus*, using the terminology of Chung and Timberlake (1985). Because the moment of utterance is contextually bound, absolute tense is a deictic system. Events described using the past tense, for example, are indicated to fall on the part of the time line that precedes the moment of utterance.

From a structural point of view, German has absolute tense and displays a past/non-past distinction, just like English and most other European languages; that is, verbs in the past tense appear with forms morphologically distinct from the present tense, but the future tense is periphrastic and uses the auxiliary *werden* ('to become') in combination with the unmarked infinitive of the verb.

The German tense system is very similar to English's, except that German lacks a commonly used progressive form. German tenses include the present, preterite (simple past), and future, and the present perfect, pluperfect (past perfect), and future perfect. The first three simple tenses locate times relative to the moment of utterance; the last three perfect tenses pick out locations on the time line using two separate points (the moment of utterance, and what Reichenbach 1947 calls the *reference time*). This can be illustrated using the English present perfect:

(2.98) Tom had already left.

Here, the *event time*, when Tom was leaving, is located before some other unspecified event that occurred in the past; the time of this other event is the reference time. Perfect tenses can usually be modified with the adverbial 'already', as is done here; indeed, some languages construct their perfect tenses solely using this adverb.

Klein (1994) presents a model of time in language that integrates both tense and grammatical aspect (section 2.4.1), by distinguishing the notion of *topic time* – the 'time span to which the speaker's claim on this occasion is confined' (p 4) – from the time span over which the situation actually holds (his *time of situation*). He argues that tense establishes a relation between the time of utterance and this topic time, whereas grammatical aspect concerns the relation between topic time and the time of situation.

NLP methods sometimes make use of tense information for lexical semantics, but this is usually either for completeness, or due to taking tense as a proxy or indicator for some other semantic category.

For example, Joanis, Stevenson and James (2008) have a feature to capture tense information as a part of their project on automatic acquisition of lexical information about English verbs. This is because some diathesis alternations in English may correlate with tense; for example, the middle construction (example 2.27) usually appears in the present tense. However, the middle is constructed differently in German, using periphrasis, and does not require tense information to be detected. As we have already seen, the perfect tense and simple past tense in English connote perfective and imperfective aspect, respectively; however, the same is not true for German.

2.5.2 *Mood and modality*

Modality is a semantic category that allows speakers to qualify what they are saying. This category may be used to indicate properties about the actuality, objective factual status, and believablity of an utterance, and it may also convey the speaker's subjective attitudes towards the utterance.

The fundamental categorisation undertaken by modality is the distinction between *realis* (actual, relating to the real world) and *irrealis* (potential or possible, relating to a hypothetical world). For example, in:

(2.99) Apparently, Rafa told his girlfriend to hide the papers.

the adverb 'apparently' is used to qualify the factual status of the proposition, with the result that the sentence does not straightforwardly assert a particular state of affairs, but leaves open the possibility that the proposition may be incorrect.

There are two major uses of modality in language: *Epistemic modality* reflects the speaker's judgement of a given proposition, as in example 2.99. This is a kind of *propositional modality*, which expresses qualities of the proposition; another kind of propositional modality is evidential modality, which can indicate the source of the proposition. *Deontic modality* expresses the speaker's attitude towards a possible future event, and is a kind of *event modality*, referring 'to events that are not actualized, events that have not taken place but are merely potential' (Palmer, 2001, p 70). In:

(2.100) (a) You may borrow my drill.

(b) You should tell me before you come over.

the speaker gives permission, or communicates an obligation, to the hearer to perform a future event.

Modality is expressed grammatically by *mood*, a grammatical system marked on the verb, and also by *modal systems* of verbs; it is also expressed lexically by adverbials (e. g., with 'perhaps', 'possibly', 'probably', etc.). German uses all of these methods.

Moods in German include the indicative, imperative (a kind of deontic modality), and the subjunctive; the subjunctive in German comes in two varieties: the Konjunktiv II (example 2.101), used for hypothetical or conditional situations; and the Konjunktiv I (example 2.102), used for indirect (reported) speech (a kind of evidential modality), as well as for wishes (the *optative mood*, example 2.103).

(2.101)
Wäre Arbeiten leicht, tät 's der Bürgermeister selber.
Is.KONJ2 working easy, do.KONJ2 it the mayor self.
'If working were easy, the mayor would do it himself.'

(2.102) Franz sagte, er habe bis 2 Uhr gearbeitet.
Franz said, he has.KONJ1 until 2 o'clock worked.
'Franz said that he had worked until 2 o'clock.'

(2.103) Lang lebe der König!
Long live.KONJ1 the king!
'Long live the King!'

German modal verbs are very similar to English ones. *Müssen* ('must') expresses deduction epistemically and obligation deontically; *mögen* ('may') and *können* ('can') express possibility epistemically, and *dürfen* ('may') and *können* express permission deontically. As with English, the use of the same forms can create ambiguity between epistemic and deontic readings:

(2.104) Er muss in seinem Büro sein.
He must in his office be.
'He must be in his office.'

This example can be taken to mean that epistemically he almost certainly is in his office, or deontically that he is required to be in his office. *Sollen* ('shall') and *wollen* ('will') are used evidentially to express reported speech (respectively, hearsay, and self-proclamations)[9]:

(2.105) Er soll steinreich sein.
He sollen.3SG.PRES.IND very rich be.
'He is said to be extremely rich.'

(2.106) Er will eine Mosquito abgeschossen haben.
He wollen.3SG.PRES.IND a Mosquito shot down have.
'He claims to have shot down a Mosquito (plane).'

9 Examples from Hammer (1983, pp 231f).

As in English, where the past tense forms of the modal verbs can be used to soften modal judgements (compare 'You may be wrong' with 'You might be wrong'), German modal verbs can be attenuated in strength by inflecting them for the Konjunktiv II mood.

Modality is sometimes analysed in terms of deixis; here, the deictic reference point is taken to be the real world. Under epistemic modality, the speaker compares the compatibility of a hypothetical possible world to the actual real one; good matches result in assumptive or deductive modals, whereas less good matches give rise to speculative modals. In deontic modality, the expressed world is an ideal moral or legal situation, and the goodness of the match is reflected in the strength of the obligation expressed.

As with tense, while modality tends to be expressed on or through verbs, it is a category that concerns the status of an entire proposition.

2.5.3 *Summary*

As discussed, both the categories of tense and modality express properties of propositions, although in German they are predominantly marked on or using verbs. As a result, these categories have much less to do with the internal semantic structure of verbs than other semantic categories such as subcategorisation and selectional preferences, and lexical aspectual structure. While I will not try to argue that we cannot expect to find lexically conditioned patterns in terms of the tenses or moods (or government by modal verbs) that a particular verb will tend to manifest with, I would argue that these patterns represent discourse or pragmatic effects more strongly than semantic ones. Because we do not expect the behaviour of verbs with respect to tense and modality to be strongly tied to their internal meaning structure, it would seem that any lexical idiosyncrasies that are observed are likely to be unhelpful to a better understanding of the verb's meaning, and should not be predictive of other semantic behaviours of the same verb.

3 LINGUISTIC RESOURCES

The empirical studies presented later in this thesis will require a source of relevant language showing a variety of verbs and verb usages, complete with some kind of syntactic and morphological analysis, whether manually-created or automatically produced; and there will need to be a sufficient quantity of this data to attain reasonably accurate estimates of verb behaviour. To this end, this chapter takes account of available linguistic resources, existing NLP tools, and applicable techniques.

Section 3.1 begins by constructing the most important resource used in this thesis, a large corpus of German with automatically-generated syntactic and morphological annotations. This corpus is made up of text drawn from a broad spectrum of language domains on a mixture of topics, and includes a large assortment of verb instances; its ample size means that it can be used to derive accurate estimates of the syntactic and morphological behaviour of German verbs. Section 3.1 also reviews the major treebanks and parsers for German. Section 3.2 introduces word vector and word embedding models, two distributional semantic techniques that I will make use of in this thesis. Section 3.3 describes GermaNet, a machine-readable lexicon, and section 3.4 discusses semantic role labelling and the German SALSA corpus. Section 3.5 is about VerbNet, a lexicon of English verbs; there is currently no German equivalent to this resource, but it is highly relevant to the work presented in later chapters. Finally, the last two sections describe practical approaches that will be central devices used in the rest of my work. Section 3.6 gives a brief outline of clustering, and section 3.7 describes a general method to evaluate the performance of an automatic system on a particular task, taken from the field of information retrieval.

3.1 CORPUS

3.1.1 *Treebanks*

A *treebank* is a corpus of text that annotates the syntactic structure of its sentences (a parse). The words in treebanks are commonly also labelled with other syntactic information, such as part of speech (POS) tags, or other morphological features (indicating, e. g., case, number, mode, etc.). Syntactic structure is modelled according to various syntactic theories; this consideration will be covered in more detail in sections 3.1.3 and 3.1.4 below. The construction of large manually an-

notated treebanks in the 1990s was instrumental in the development of reliable statistical parsers.

The first large-scale treebank was the Penn TreeBank (PTB: Marcus, Santorini and Marcinkiewicz, 1993), which contained phrase structure analyses of English. It comprises 2,499 articles from the Wall Street Journal (WSJ), totalling over one million words.

The NEGRA (Skut et al., 1997) project was an early treebank of German, consisting of 20K sentences from the *Frankfurter Rundschau* (355K words) and covering a variety of domains. NEGRA also annotates phrase structure like the PTB, but allows for branches to cross in the syntactic analyses, allowing non-local dependencies to be encoded without using traces. POS is indicated for individual tokens using the Stuttgart-Tübingen tag set (STTS: Schiller, Teufel and Thielen, 1999).

Further, NEGRA includes some dependency grammar information by using *edge labels* to record the syntactic function of a constituent (a list of these is reproduced in table 5.1). Because the word order of German is relatively free, syntactic analysis must indicate the syntactic function of constituents, since this cannot be derived from word order information the way it largely is in English (cf. figure 3.1).

The TIGER (Brants et al., 2002) project continued on where NEGRA left off, annotating another 50K sentences from the *Frankfurter Rundschau* (900K words). The NEGRA annotation scheme was extended for TIGER to include morphological features and *lemma* information (the uninflected base form of a word, such as the infinitive for verbs) about the terminal nodes of the tree; other changes made included:

- better marking of expletive *es* 'it';
- better marking of named entities using the `PN` (proper name) label; and
- more functionally descriptive labels for PPs instead of just `MO` (modifier), allowing prepositional objects (`OP`) and collocational verb constructions (`CVC`) to be distinguished, which reflects the argument/adjunct distinction (cf. section 2.3.3) for PPs.

The third German treebank is the TüBa-D/Z (Telljohann, Hinrichs and Kübler, 2004), consisting of 105K sentences (1960K words) taken from 3,800 articles from *die tageszeitung* (taz) newspaper. Similarly to TIGER, TüBa-D/Z annotates phrase structure and records syntactic functions as edge labels, indicating the head/non-head distinction; it also includes STTS POS tags, morphological information, and lemma information about individual tokens. Beyond this, the corpus annotates layers covering named entities, anaphora, and coreference, as well as some word sense tags for individual tokens. The annotation scheme used does not allow for crossing branches, and, compared to TIGER, TüBa-D/Z has relatively deep tree structures. A more in-depth

comparison of the TIGER and TüBa-D/Z treebank projects is presented by Dipper and Kübler (2017). In the last ten years, NLP work on the training and evaluation of German parsers has tended to focus on the TIGER corpus, and this thesis will tend to do so as well.

3.1.2 *TreeTagger: The German part of speech tagger*

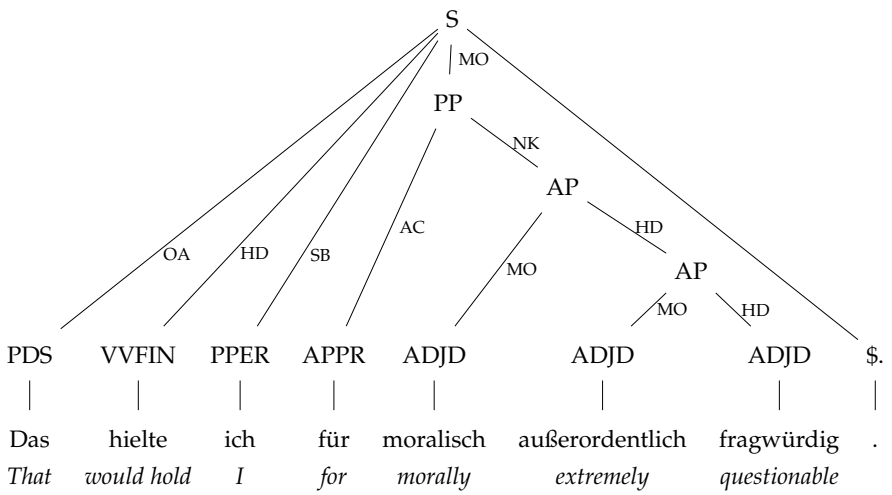
A part of speech tagger takes in raw text and automatically assigns to each word a part of speech tag (i. e., noun, verb, adjective, adverb, determiner, etc.). The TreeTagger (Schmid, 1994, 1995) is a Markov model tagger that makes use of a decision tree to estimate transition probabilities between hidden states. It achieves 97.5% accuracy on a corpus of newspaper text from the *Stuttgarter Zeitung* (20K training tokens, 5K testing tokens). The TreeTagger can also lemmatise German text.

3.1.3 *Constituency parsers*

The task of a parser is to construct the syntactic tree associated with a grammatical analysis of a sentence, given the words of the sentence and their respective POS. Constituency parsers generate syntactic trees described by context free grammars (CFGs), whereby a non-terminal node (such as S, for sentence) is replaced, using a grammar rule, by one or more terminal or non-terminal nodes (such as $S \rightarrow NP\ VP$, for a noun phrase followed by a verb phrase). This is the grammatical formalism used in the phrase structure analyses of early and influential treebanks such as the PTB and TIGER. Diagramming these rule applications in the vertical axis results in tree structures. Figure 3.1 shows a constituency parse of the sentence 'I would consider that morally extremely questionable.'

The availability of manually-annotated syntactic treebanks allowed for the rapid development and iterative improvement of robust automatic parsers. Statistical parsers of phrase structure grammars typically model the language using a PCFG. Under this view, a single non-terminal may be rewritten by multiple incompatible grammar rules, making it syntactically ambiguous; the PCFG guarantees that in such cases the *probabilities* of the applicable grammar rules will sum to one.

Treebanks of a sufficient size allowed constituency parsers to be trained and evaluated in a *supervised learning paradigm*. This is accomplished by transforming the treebank into a set of labelled data points. Thus, the set of locations in a syntactic tree become a collection of input data, whose members are described by sets of features (for instance, the POS tag of the current tree node, the current height in the tree, the token associated with the leftmost daughter node, etc.). Each of these input data points is associated with a ground truth output,




```
(S (OAPDS Das) (HDVVFIN hielte) (SBPPER ich)
(MOPP (ACAPPR für) (NKAP (MOADJD moralisch)
(HDAP (MOADJD außerordentlich) (HDADJD fragwürdig))))
($. .))
```

Figure 3.1: A phrase structure grammar analysis of a sentence in the TIGER corpus, showing how edge labels indicate syntactic function. The figure also shows a *bracketed representation* of the same structure.

i. e. the decision to *reduce* a set of nodes to a non-terminal using a particular PCFG rule, versus the decision to *shift* the current node onto a stack to permit the parser to reduce it at a later time.

In short, the treebank used to train the parser attests to the set of legal PCFG rules for the language, and the sample of tree structures contained in the treebank allow statistical systems to estimate the quality of a particular syntactic analysis of a particular local syntactic neighbourhood. This ability to estimate goodness can then be used to search for the best possible syntactic analysis for a whole sentence.

The Stanford Parser (Klein and Manning, 2003) is an unlexicalised PCFG parser.[1] Rafferty and Manning (2008) developed and made available a model for parsing German, trained on TIGER.

The Berkeley Parser (Petrov et al., 2006) is another unlexicalised PCFG parser that incorporates smoothing and other tricks on top of the techniques used in the Stanford Parser. Petrov and Klein (2007) produced and distributed a German parsing model for the Berkeley Parser trained on TIGER.

1 In *lexicalised parsers*, developed by Charniak (1997, 2000) and Collins (2003) among others (Bikel, 2004; Charniak and Johnson, 2005), phrasal nodes are annotated with their syntactic heads, effectively meaning that the syntactic head of a phrase predicts (or 'projects') the other elements of its constituent, which accords with the view often taken in syntactic theory.

	TIGER	TüBa-D/Z
Berkeley Parser	69.81	83.97
Stanford Parser	58.07	79.24

Table 3.1: Labelled F_1 scores for constituency parsers reported by Kübler (2008).

Constituency parser performance is often measured using the PARSEVAL metric (Black et al., 1991), which counts parentheses in a bracketed representation of the sentence: A bracket in the system output which is in the same position in the gold standard is a 'correct' answer, and one which has no corresponding bracket in the gold standard is a 'false' answer. That is, each constituent in the parse output that exactly matches a constituent in the gold analysis, having the same span and non-terminal label, is counted as a correct answer. Given the confusion matrix, the measures precision and recall can be computed and summarised with an F_1 score (cf. section 3.7).

Kübler (2008) organised the PaGe 2008 shared task to directly compare the performance of German constituency parsers on both TIGER and TüBa-D/Z. Training data were 20,894 sentences from each of the two treebanks; development and test sets were 2,611 sentences each. The PARSEVAL metric was used to compare the Berkeley and Stanford parsers, with Berkeley delivering the best performance as shown in table 3.1.

3.1.4 *Dependency parsers*

Dependency grammar describes syntactic structure as a series of directed links, each emanating from a *head word* and terminating at that word's *dependent*; the main verb of the sentence is the root of the tree structure, and all other words in the sentence are either directly or indirectly dependent on this root node. Dependency grammars are less explicitly concerned with the word order of sentences, compared to phrase structure grammars; this can make them easier to use for languages with a freer word order than English, such as German.

Development of German dependency parsers has been based on automatic conversions of the TIGER treebank into a dependency treebank format; the most recent and detailed of these conversions was done by Seeker et al. (2010).

Figure 3.2 shows the dependency grammar counterpart of the example sentence from figure 3.1. The figure makes clear that the edge labels from the TIGER corpus can be applied without modification to the new structure of the sentence, and are well suited to indicate the syntactic function of dependency links.

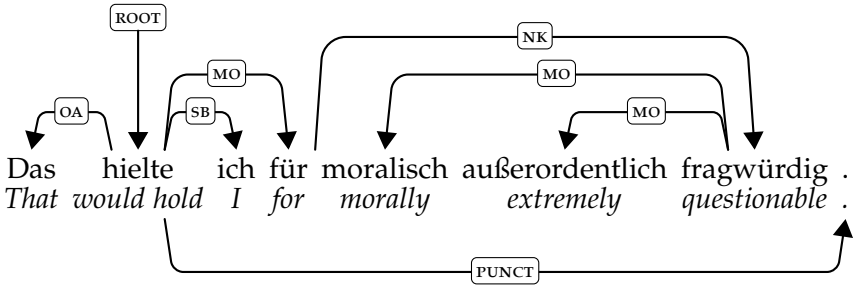


Figure 3.2: A dependency grammar analysis of a sentence in the TIGER corpus. 'I'd consider that morally extremely questionable.'

Dependency parsers can be evaluated using their labelled attachment score (LAS: Buchholz and Marsi, 2006), defined as the percentage of tokens (not including punctuation) for which the parser is able to predict the correct head and dependency label.

One early stadium for statistical German dependency parsing was the CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi, 2006); in addition, the 2008 PaGe shared task (Kübler, 2008) also evaluated a single dependency parser. The CoNLL 2009 shared task was mostly focused on semantic role labelling (cf. section 3.4), but also included a dependency parsing evaluation setup to measure the performance of pipelined labelling systems.

These competitions were dominated by two influential parser projects: the MaltParser (Hall and Nivre, 2008), which popularised the transition-based approach to dependency parsing; and the MST Parser (McDonald et al., 2005), which was a prominent example of the graph-based method.

Bohnet et al. (2013) presented the `mate-tools` parser, which performed joint labelling of morphological features, part of speech, and dependency structure, with a specific focus on better performance for richly inflected languages such as Czech, Finnish, German, Hungarian, and Russian. The parser also performs lemmatisation. The German model made use of the SMOR morphological analyser (Schmid, Fitschen and Heid, 2004), as well as 800 Brown clusters (Brown et al., 1992) derived from the Huge German Corpus (204M tokens)[2]. The `mate-tools` parser recorded state of the art results, as measured on the test sets produced by automatic TIGER conversions as performed by CoNLL 2009, and by Seeker et al. (2010). The winning German parser on the Shared Task on Parsing Morphologically Rich Languages (Seddah et al., 2013) later in the same year also used the `mate-tools` parser as part of an ensemble.

2 `https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc/`.

3.1.5 *SdeWaC*

Data-based NLP methods have relied on access to large quantities of text since the beginning. The Brown Corpus (Kučera and Francis, 1967), with approximately one million words of English, was widely used in early work. As more accurate models can be derived from more data, there followed a series of corpora of increasing size, such as the British National Corpus (100M words, Burnard, 2000).

An advantage of corpus linguistics is that the language samples under investigation can be said to be representative, in some way, of actual language use by real human beings. This is reflected in the care often taken to balance corpora for genre or domain, historical period, written or spoken language, fact or fiction, and so on.

In the past 20 years, NLP research has turned to the World Wide Web as a large source of text (Kilgarriff and Grefenstette, 2003). While wild text harvested from the Internet at random is not manually balanced, it is also in some way reflective of real language use, and covers a range of domains. In a world where bigger is better, the idea of Web as Corpus (WaC) is hard to argue with; German makes up about 2.7% of the web, according to a recent survey (W3Techs.com, 2020).

One resource constructed in this vein is deWaC (Baroni et al., 2009), comprising 10^9 words of German extracted from Web search results for random combinations of search terms; the text was automatically POS-tagged and lemmatised by the TreeTagger (section 3.1.2). Faaß and Eckart (2013) created the SdeWaC corpus by filtering deWaC in an effort to reduce noise. First they removed duplicate sentences; after this, they used a rule-based dependency parser to scores sentences with a per-token error rate, and removed sentences with particularly low parsabillity scores. The remaining body of text comprises 880M words in 45M sentences.

I used the CLOU (Cluster of UNIX Machines) compute cluster of the Humboldt University to parse the entirety of SdeWaC with the `mate-tools` parser. This took about 4,360 hours of compute time in total, which was accomplished over the span of about five weeks; figure 3.3 shows this progress. The automatically parsed SdeWaC is used throughout this thesis as a source of linguistic data on verb behaviour. In future chapters I will use the term SdeWaC to refer to the treebank automatically produced by the `mate-tools` parser. As we will see in chapter 5, the drawback that automatic parses are less accurate than manual annotations is outweighed by the benefit of the larger volume of training data and greater vocabulary.

3.2 WORD VECTORS

A vector space model or word space model (WSM) (Sahlgren, 2006; Turney and Pantel, 2010) is a model of lexical semantic similarity

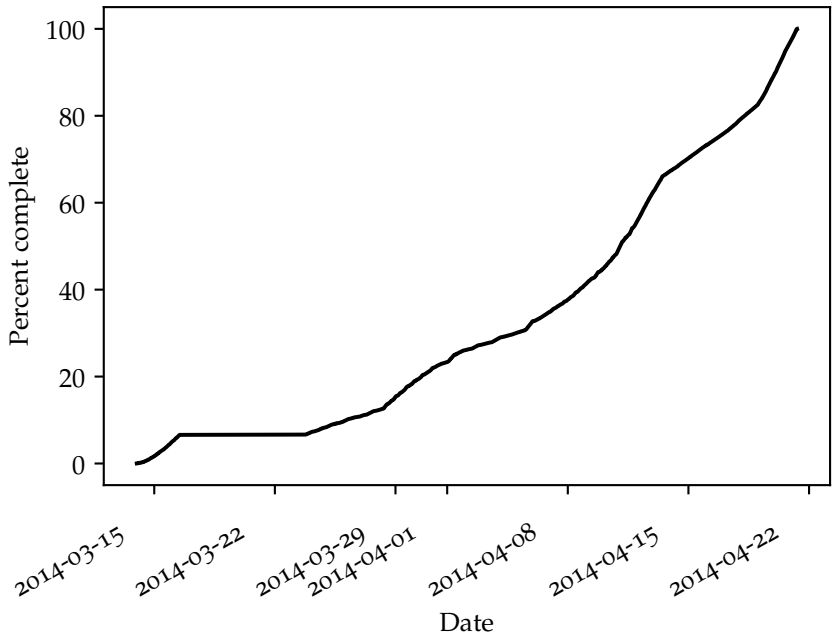


Figure 3.3: Timeline of automatically parsing the SdeWaC corpus.

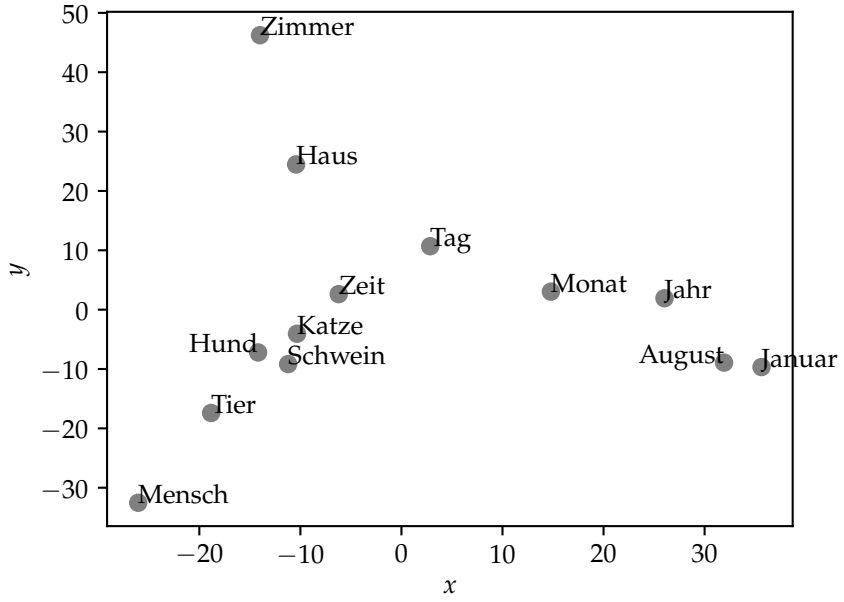


Figure 3.4: Illustration of a word space model. The word vectors have been projected to a two dimensional space using a PCA.

that operates using a spatial metaphor of word meaning. A useful feature of most WSM models is that they can be constructed in an unsupervised manner from raw text, with no morphological or syntactic analysis necessary. As such, these models have long been used in information retrieval, where indexing and searching large volumes of unstructured text is a common objective.

Each word in the vocabulary is associated with a n-dimensional vector, representing a point in a space. Figure 3.4 shows an example word space with two dimensions, but usually much higher dimensions are used.[3] The principal intuition underlying this setup is that words that are semantically related to each other are placed close together in this space, and words that are unrelated will tend to be far apart from each other. In the figure, this structure is reflected in the fact that time expressions such as *Tag* 'day', *Monat* 'month', *Jahr* 'year', and the months *Januar* and *August* are placed together in one part of the space; and also that words for animate entities are similarly clustered together: *Katze* 'cat', *Schwein* 'pig', *Hund* 'dog', *Tier* 'animal', *Mensch* 'human'. A problem common to many word vector approaches arises from the fact that, typically, only one vector is constructed per lexeme, so that different senses of the same word will be conflated using these methods.

The next ingredient for a distributional semantic model is to derive the vector for a given word from statistics about that word's distribution. This follows the *distributional hypothesis* (cf. chapter 1), so that words with similar distributional properties are assigned similar word vectors, and, as such, can be said to have similar meanings. For a more in-depth review of distributional semantics techniques and theory, the reader is referred to Manning and Schütze (1999).

An early recipe satisfying both these criteria was the vector-space model (Salton and McGill, 1983) from 1960s. This relies on a word-document matrix, wherein each row represents a document in a collection to be indexed, and each column represents a word in the vocabulary. The entry on the i^{th} row and j^{th} column will be exactly the number of times word j appears in document i. A row of this matrix is a *bag of words* representation of a single document (a bag is like a set, in that it is an unordered collection, but, unlike a set, each member may belong to a bag more than one time). The column associated with a particular word can be termed a *context vector*, showing the kinds of text a word can appear in.

Count entries may be transformed by some weighting scheme to reduce the impact of uninformative words on similarity scores. An example is tf-idf (*term frequency–inverse document frequency*), where the *term frequency* (Luhn, 1957) for a given word-document pair increases

3 Actually, the word vectors shown in figure 3.4 have 50,000 dimensions, but a principal components analysis (PCA) transform has been used to find the two dimensional basis vectors that represent the directions of greatest variance, and the vectors have been projected down to this basis for ease of illustration.

linearly as the word appears more often in the document; this is multiplied by an inverse *document frequency* (Spärck Jones, 1972), which measures how many other documents the given word appears in (indicating if the word is very general and common, or more likely to be specific to a particular topic or domain).

The similarity of two words can be calculated in a number of ways, such as the Euclidean distance separating the two vectors, but is usually obtained as the *normalised vector similarity* (the dot product of the two vectors when they are normalised to have unit length, representing the cosine of the angle separating the vectors).

Word-document matrices are often very large; Webster's dictionary of English lists 470,000 different head words, for example, and the documents available to be searched could easily number in the millions. Simultaneously, the matrix will inevitably be very sparse, with only a small number of nonzero entries; this is a result of the Zipfian nature of language (Zipf, 1949), whereby the most frequent word types in a language disproportionately represent the vast majority of word tokens in a given sample of that language.

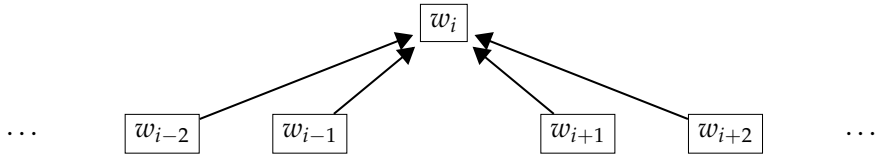
Rank-reduction methods such as singular value decomposition (SVD) can be applied to the co-occurrence matrix, creating an approximation to the information contained in the original matrix, and allowing word vectors to be expressed using fewer dimensions. This effects a kind of regularisation, pushing the vectors of words that occur in similar contexts to more closely resemble each other. This was the primary innovation behind Latent Semantic Analysis (Landauer and Dumais, 1997), a popular technique for information retrieval in the 1980s and 1990s.

Word space models (Schütze, 1993; Sahlgren, 2006; Turney and Pantel, 2010) are a more fine-grained spin on the word-document approach: Instead of documents, we consider some other smaller lexical context, such as a sentence or, smaller yet, the words in a window of fixed size. For every pair of words inside one of these contexts, the word-word co-occurrence count for that pair is incremented by one. This is repeated for all lexical contexts in a large text corpus. The result of this procedure is a word-word matrix, where each column and each row represents one word of the vocabulary. The word vector for a given word is then just the row or column corresponding to that word; the matrix containing co-occurrence counts is symmetric, so the choice of columns or rows does not matter.

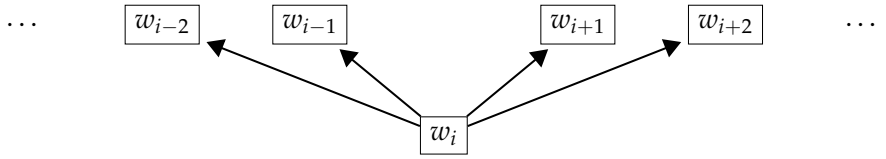
I create and make use of a word space model of German in section 6.2.3 for estimating the semantic similarity of German nouns.

3.2.1 *Word embeddings*

Word embeddings are a more recent development that create dense vector representations for words, typically with a low number of di-



(a) A continuous bag of words (CBOW) neural network attempts to predict a word from its linguistic context.



(b) A skip-gram neural network attempts to predict a context from a word.

Figure 3.5: Word embedding model architectures.

mensions (several hundred). As in word-space models, a linguistic context is formalised as a fixed window around a central target word. A neural network is then trained to predict the target word from its context (the continuous bag of words model, or CBOW); alternatively, the network can be used to predict the context from the target word (the skip-gram model). These two schemes are diagrammed in figure 3.5. This prediction task, a kind of supervised learning problem, is accomplished on the basis of unannotated textual input, similarly to other word vector models. As a result of the training, the network learns representations of the words in the vocabulary that maximise the likelihood of predicting the contexts in which that word is observed to occur in the training corpus. Much like in other word vector approaches, words that occur in similar contexts are induced to have similar vector representations.

Use of word embeddings became widespread in NLP research following the release of efficient tools for creating them, such as `word2vec` (Mikolov et al., 2013), GloVe (Pennington, Socher and Manning, 2014), and C&W (Collobert and Weston, 2008). Embeddings have been demonstrated to be an effective knowledge source for multiple NLP tasks (e. g., Collobert et al., 2011), and the lexical information contained in word embeddings has produced performance improvements in NLP applications such as named entity recognition (NER), POS tagging, semantic role labelling (SRL, Collobert et al., 2011), sentiment analysis (Kim, 2014; Iyyer et al., 2015; Tai, Socher and Manning, 2015; Looks et al., 2017; Yu et al., 2017; Cliche, 2017), automatic summarisation (Paulus, Xiong and Socher, 2017), machine translation (MT, Bahdanau, Cho and Bengio, 2015; Sennrich, Haddow and Birch, 2016; Bojar et al., 2016), and parsing (Chen and Manning, 2014; Dyer et al., 2015; Straka et al., 2015). Word embeddings remain an active field of research, and are the dominant approach to applied distributional semantics today;

increasingly large and sophisticated models such as BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020) routinely make popular science headlines.

I create and make use of German word embeddings as features to a supervised classifier in section 8.1.8.

3.3 GERMANET

The WordNet project (Fellbaum, 1998), started in the 1980s at Princeton University, is a machine-readable database of English that is structured both like a thesaurus, by arranging the lexicon according to the semantic relation of synonymy, and also like an ontology, through the systematic marking of *hyponymy* relations ('is a' or 'kind of' relations). The GermaNet project (Hamp and Feldweg, 1997), developed at the University of Tübingen, can be considered a relatively faithful translation of WordNet into German.

Each head word listed in the dictionary is associated with one or more word senses, each having an associated part of speech. Each of these word senses belongs to a *synset*, a set of synonymous word senses that all have the same meaning; every synset has a definition (or *gloss*) and a set of ontological relations.[4] GermaNet lists 84,882 noun, 14,331 verb, and 12,148 adjective senses; it contains 84,584 synsets, each of which contains on average 1.31 member word senses.

Figure 3.6 shows the word senses listed under the head word *sagen* 'to say'. In the diagram, there are three synsets, each of which collects together word senses with the same meaning.

The synset of the first sense listed, perhaps the most general use of the word, has diverse hyponyms including `empfehlen v 2` ('to recommend something'), `verraten v 3` ('to betray a confidence') and `fragen v 2` ('to ask a question'). Its hypernym is `verbal kommunizieren v 1` ('to communicate verbally'), which in turn has a hypernym of `kommunizieren v 1` ('to communicate'), which is followed by `interagieren v 1` ('to interact') and `handeln v 5` ('to do something'), before reaching the GermaNet root node `GNROOT`.

Hyponymy organises the GermaNet dictionary into a directed acyclic graph structure, which can be viewed as a kind of ontology over concepts. A fragment of the hyponymy tree structure of GermaNet is pictured in figure 6.5 on page 114.

In addition to hyponymy, WordNet and GermaNet also list other kinds of semantic relations:

- holonymy-meronymy relations ('part of' or 'has part' relations; `Schmuckstück n 2` 'piece of jewellery' has a component mer-

4 To be precise, in WordNet, synsets have glosses, whereas, in GermaNet, word senses have glosses. In both, semantic relations may hold between either synsets or word senses.

1. • `sagen v 1`: *bestimmte Worte sprechen, mit direkter oder indirekter Rede verwendet* ('to speak certain words, used with direct or indirect speech');
   • `äußern v 2`: *etwas äußern, mitteilen* ('to express something, communicate');
   • `meinen v 5`: *etwas sagen oder etwas aussagen* ('to say something or state something');
2. • `sagen v 2`: *mit Dativ, Worte an eine bestimmte Person oder Personenkreis richten* ('with the dative, to address words to a specific person or group of people');
3. • `sagen v 3`;
   • `bedeuten v 1`: *unpersönlich, etwas ankündigen* ('impersonal, to announce something');
   • `heißen v 2`: *einen bestimmten Sinn haben* ('to have a certain meaning').

Figure 3.6: GermaNet entries for the head word *sagen* 'say', showing the synset structure of the dictionary.

onym of `Schmuckstein n 1` 'gemstone' and a substance meronym of `Edelmetall n 1` 'noble metal');

- antonymy (`willkommen j 1` 'welcome' is the opposite of `unerwünscht j 1` 'unwanted');
- causation (`heften v 1` 'to staple' causes `haften v 3` 'to adhere');
- entailment (`gelingen v 1` 'to succeed' entails `versuchen v 3` 'to try'); as well as
- more general relationships (`Venezuela n 1` is related to `Bolivar n 1`).

We shall meet GermaNet again in chapter 6.

3.4 SEMANTIC ROLE LABELLING AND SALSA

Semantic role labelling (SRL) is the task of classifying syntactic arguments to a particular verb according to their semantic role. The labelled categories can be thematic roles such as Agent and Patient (section 2.3.4), FrameNet frame-roles, or PropBank-style numbered arguments. The first automatic SRL system was first demonstrated by Gildea and Jurafsky (2002), following the release of FrameNet. A survey of the SRL field is given by Màrquez et al. (2008).

The first major project in SRL was FrameNet (Baker, Fillmore and Lowe, 1998), started by Charles Fillmore, and built on his theory of frame semantics (Fillmore, 1976). Semantic frames are schematic descriptions of a situation, usually specific to a particular category of situations (as in SMUGGLING, with its roles of PERPETRATOR and GOODS). Frames are associated with a list of lexical units (both verbs and nouns) that they can be realised by; COMPLAINING, example, can be evoked by 'belly-ache', 'gripe', 'grumble', and so forth. Frames combine valency information with *frame elements* (semantic roles). An example is the CONQUERING frame ('capture', 'conquer', 'fall', etc.), which includes the *core participants* (arguments) of the CONQUEROR, and a THEME that is conquered. *Non-core participants* (adjuncts) to the frame can be DEGREE, INSTRUMENT, MANNER, and so on. FrameNet frames are also associated with annotated sentences, manually contrived to illustrate the lexical realisations of frames, which show sentence structure and combinations of syntactic arguments. The FrameNet resource includes more than 200,000 manually annotated sentences illustrating more than 1,200 semantic frames, and lists around 13,000 word senses.

PropBank (Palmer, Gildea and Kingsbury, 2005) takes a different approach to semantic roles. It annotates all verbs in the PTB, producing a corpus more representative of actual language use than FrameNet's. Furthermore, the semantic roles used in PropBank are much less intricate than those in FrameNet. Each verb in PropBank has a *frameset* showing possible combinations of obligatory arguments; the arguments are not named, but rather numbered, starting from zero. Generally, Arg0 and Arg1 are used systematically as proto-Agent and proto-Patient roles, respectively, but 'no consistent generalizations can be made across verbs for the higher-numbered arguments' (Palmer, Gildea and Kingsbury, 2005, p 75). Adjuncts are given named roles, such as ArgM-LOC for locatives, ArgM-MNR for manner adverbials, and ArgM-TMP for temporal adverbials.

The SALSA project (Burchardt et al., 2006; Rehbein et al., 2012) provides manual semantic role annotations on the TIGER corpus, based on the English FrameNet, and adding new German frames where necessary. German verb and noun predicate senses are numbered as necessary. The corpus annotates around 20,000 verbal and 17,000 nominal instances, using 1,950 different frames. The SALSA annotations were converted semi-automatically to PropBank-style annotations for the CoNLL 2009 shared task on syntactic and semantic dependency labelling (Hajič et al., 2009).

I will make use of SALSA and the CoNLL 2009 shared task in section 8.2.

3.5 VERBNET

VerbNet (Kipper-Schuler, 2005) is a hierarchical database of English verbs that contains syntactic and semantic information. It is based on Levin's classification of verbs according to alternation behaviour (Levin, 1993, cf. section 2.3.5), although it has been extended several times with new verb classes (Korhonen and Briscoe, 2004; Kipper et al., 2006). Not only are verbs arranged into classes in VerbNet, but the resource also records their syntactic and semantic properties. VerbNet has 23 thematic roles that are valid across all verbs, which are familiar from linguistic theory: Agent, Patient, Instrument, Theme, Recipient, etc. Of course, VerbNet lists argument realisation patterns (subcategorisation frames) for verbs; it also has information on selectional restrictions in the form of semantic predicates which must hold of a potential argument to the verb, for example concrete, animate, plant, and idea. The latest version of VerbNet (version 3.2)[5] has 273 top-level classes and 214 sub-classes; it lists 8,537 different verbs.

The combination of thematic role and syntactic frame information makes VerbNet potentially useful for training SRL systems, although, in practice, PropBank has so far produced better results (Zapirain, Agirre and Màrquez, 2008; Merlo and Van Der Plas, 2009).

The SemLink project (Loper, Yi and Palmer, 2007)[6] has manually created mappings between VerbNet classes, PropBank rolesets, and FrameNet frames.

There is no VerbNet for German yet, although there have been some efforts in this direction. For example, the SR3de project (Mújdricza-Maydt et al., 2016; Hartmann et al., 2017)[7] has mapped annotations for 3,000 verb predicate instances from the CoNLL 2009 Shared Task to VerbNet thematic roles that are translated into and slightly adapted for the German language. This means that this small corpus can be used to compare SRL using PropBank-, FrameNet- and VerbNet-style roles. In SR3de, predicates are not mapped to verb classes, but rather to their most appropriate GermaNet sense(s), which also makes this a sense-tagged corpus.

3.6 CLUSTERING

Clustering is an unsupervised data analysis technique used to discover structure inside an unlabelled data set. The purpose is to group objects into *clusters*, such that objects in the same cluster tend to be similar to each other in some way, and objects in separate clusters tend to be dissimilar. Clustering methods commonly use a *distance measure* to formalise the idea of how similar two objects under consideration

5 https://verbs.colorado.edu/verbnet/
6 http://verbs.colorado.edu/semlink/
7 https://www.cl.uni-heidelberg.de/projects/SR3de/

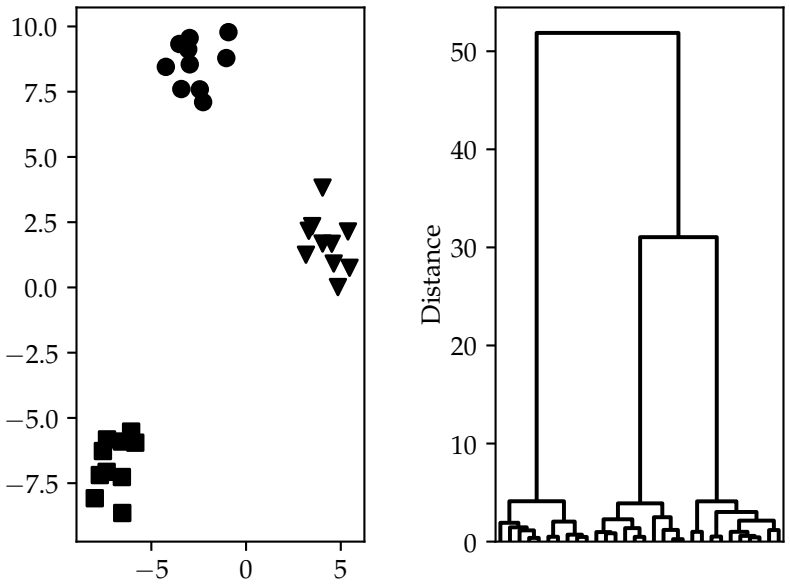


Figure 3.7: Example cluster analysis: Objects in two-dimensional space (left); dendrogram of hierarchical clustering with Ward's criterion (right).

are. Examples of distance measures include the Euclidean distance for points in space or the Hamming distance for binary strings.

Hierarchical clustering is a simple method that makes greedy use of a distance function by agglomerating objects into groups in a bottom-up manner: The algorithm begins by putting each object into its own singleton cluster of size one. At each subsequent point in time, the two clusters that are closest to each other under the distance function are merged, until all objects are collected in a single cluster. The result of this procedure is a *dendrogram* that records the history of these merges; by reading off the appropriate level of the dendrogram, hierarchical clustering can group a data set into any number of clusters desired. Various *linkage functions* are used to calculate the distance between pairs of clusters before each merge:

SINGLE LINKAGE the distance between two clusters is the smallest distance between their constituent objects;

COMPLETE LINKAGE the distance between two clusters is the largest distance between their constituent objects;

AVERAGE LINKAGE the distance between two clusters is the arithmetic mean of all distances between all of their constituent objects; equivalent to the distance between cluster centroids; and

WARD'S LINKAGE the distance between two clusters is the sum of the squares of all distances between all their constituent objects.

Figure 3.7 shows an example cluster analysis of contrived data in two dimensions. On the left, the true distribution of the data is illustrated with different shapes; on the right is a dendrogram induced with Euclidean distance as a metric, and using Ward's criterion for hierarchical clustering (Ward, Jr, 1963). The dendrogram reflects the same intuition delivered by the naked eye: The data are well separated and the structure is best represented by three clusters; two of these are slightly closer to each other than they are to the third.

The *k*-means clustering algorithm (Forgy, 1965) is an iterative method for clustering objects into a fixed number of groups, sometimes referred to using the name of the closely related Lloyd's algorithm (Lloyd, 1982). Clusters are characterised by their *centroids*, the mean value of the objects belonging to that cluster. At each step, objects are assigned to the cluster whose centroid is closest; equivalently, the cluster centroids divide the space of the data set into Voronoi cells, and objects are partitioned into groups according to which cell they fall in. After cluster membership is updated in this way, the cluster centroids are recalculated; the process of assigning and updating is then repeated until a stable solution is found. Clusters are typically initialised either by random assignment of objects to clusters, or by picking random objects as initial centroids.

Both hierarchical and *k*-means clustering are *hard clustering* techniques, which partition n samples into k clusters ($k \leq n$). There also exist *soft clustering* techniques, where samples can be associated with more than one cluster, and membership in a cluster is a gradual scale. Hard clustering will be a central motif of this dissertation, since I will use a clustering-based method for evaluating various distributional representations of verbs. Section 4.2 in the next chapter presents a review of prior work on automatic verb clustering. I will also make use of soft clustering techniques in chapter 6.

3.7 INFORMATION RETRIEVAL

Information retrieval (IR) is the task of finding relevant information inside some large collection of documents. As an applied branch of NLP with a long history, it has well-developed quantitative methods, including widely-used evaluation metrics; after all, measuring the performance of implemented systems is important to determine objectively how to improve those systems. For the purposes of this dissertation, such evaluation measures are of critical importance for judging the quality of the systems that I develop, and are also needed for comparisons to previous work by other researchers.

Many evaluation paradigms begin with a pre-defined test set, where some set of examples are associated with a desired label that an automatic system should ideally produce (the *ground truth*). For document retrieval for a particular search query, for example, this might take the

	Ground truth	
	Relevant	Irrelevant
Retrieved	TP	FP
Not retrieved	FN	TN

Table 3.2: The categories counted in a confusion matrix.

form of labelling each of a set of documents as being either relevant or irrelevant to the query. A system's observed output in practice can then be compared against the ground truth. This comparison can result in four possible outcomes, with two kinds of 'correct' answers and two types of 'incorrect' answers:

1. True positives (TP) are those documents found by the search system which are actually relevant;
2. true negatives (TN) are documents correctly not returned by the system because they are actually irrelevant;
3. false negatives (FN, 'type I errors') are documents not found by the system that are actually relevant; and
4. false positives (FP, 'type II errors') are irrelevant documents returned by the search system that actually should not be included in the search results.

The counts of these outcomes over the whole test set can be tabulated in a *confusion matrix*, depicted in table 3.2. On the strength of the confusion matrix, we can define several evaluation measures. A naïve way to measure the goodness of a system's output is its *accuracy*, the fraction of all of items that are in some way 'correct':

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

It is more common, however, to judge performance with two other statistics: precision and recall. *Precision* is the fraction of the system's output that is actually correct:

$$P = \frac{TP}{TP + FP} \tag{3.2}$$

Recall is the fraction of all correct answers that the system actually found:

$$R = \frac{TP}{TP + FN} \tag{3.3}$$

The two measures are often summarised by their harmonic mean, the *F-score*. This is because both precision and recall measure valuable attributes of a system, and also because all automated methods can be described in the limit by a particular quality of performance on a given task; at this level, twiddling the algorithm's parameters is usually only able to trade greater precision against poorer recall (or vice-versa). The *F*-score captures this shared emphasis on making high-quality predictions, and also making sufficiently many of them:

$$F_1 = \frac{2PR}{P+R} \tag{3.4}$$

The *F*-score ranges from a largest possible value of 1.0, when both precision and recall are perfect, to a least possible value of 0.0, when either precision or recall is zero. I will make use of precision, recall, and *F*-score as evaluation measures in multiple places. I will also on occasion report accuracy scores.

4

RELATED WORK

This dissertation explores four concrete domains of NLP. First, chapter 5 describes the construction of a system to automatically determine the subcategorisation frame of a given verb instance, and applies this method to the parsed SdeWaC corpus produced in section 3.1.5. This system is evaluated in various ways, including using a previously-published automatic verb classification task. Chapter 6 extends this system to include information about a verb instance's arguments, inducing models of verbal selectional preferences. Next, in chapter 7 and chapter 8, a small corpus of German verbs is manually annotated for features of lexical aspect; this resource is then used to train a set of classifiers that can be applied to arbitrary verb instances. Finally, I test the efficacy of these classifiers in several extrinsic tasks.

Naturally, the domains of applied NLP that I prospect here have already been described in previous research. Thus, to provide context to my work, this chapter reviews the literature in each of these domains. I will also try to motivate these applications and argue that they are pragmatically useful. Section 4.1 appraises previous work on subcategorisation acquisition; this leads naturally into section 4.2, which summarises research done on automatic verb classification. Section 4.3 surveys related work on modelling selectional preferences; and section 4.4 considers the field of computational aspect.

4.1 SUBCATEGORISATION ACQUISITION

Automatic subcategorisation acquisition is a task that has many applications, because models of SCF can capture the behaviour of a verb's predicate-argument structure, and this information is useful for any NLP task that focuses on the verb, such as parsing, verb clustering, SRL, machine translation (MT), and WSD. An early study is presented by Carroll, Minnen and Briscoe (1998), who modified a PCFG parser to use SCF probabilities and demonstrated a significant improvement in performance. Carroll and Fang (2004) showed that subcategorisation information could improve the coverage of an Head-Driven Phrase Structure Grammar (HPSG) parser. SRL systems described by Grenager and Manning (2006), Lang and Lapata (2010) and Titov and Klementiev (2012) also make use of SCF frequencies as features, since some kind of treatment of argument linking allows theta roles to be predicted from syntactic roles. Semantic roles have wide applications in NLP; for example, Surdeanu et al. (2003) show how an SRL system can be effectively used to perform information ex-

traction (IE). Kohomban and Lee (2005) showed the benefit of adding features representing the subcategorisation behaviour of verbs to an automatic WSD system. My Master's thesis (Roberts, 2011) also explored the use of SCF as a knowledge source for WSD.

As such, it is not surprising that this is a task with a relatively long history for the field of NLP. Thus, a review of the prior work in this area covers several different epochs of language processing techniques.

The story begins with work in English by Brent (1991, 1993), who collected a small set of SCFs (five and six, respectively) from completely unprocessed text, using manually written syntactic cues to represent particular context patterns, often involving closed-class lexical items such as pronouns or proper names. This work had to confront an essential challenge to automatic subcategorisation acquisition, which is that automatic analysis produces occasional errors, which lead to noisy data. In response, Brent developed the technique of *hypothesis testing* to statistically filter automatically-generated SCF observations; Brent's method, which continues to be widely used, is based on the binomial hypothesis test (BHT), which can estimate the likelihood that a verb is incorrectly tagged with a particular SCF.

Manning (1993) continues in this vein, using a chunk parser and BHT filtering to collect information on 19 SCFs; evaluation was done by hand against the Oxford Advanced Learner's Dictionary (Hornby, 1974).

On a larger scale, Briscoe and Carroll (1997) used a statistical parser to collect counts of 163 SCFs, defined by the union of the subcategorisation frames used in the ANLT (Boguraev et al., 1987) and COMLEX (Grishman, Macleod and Meyers, 1994) machine-readable dictionaries. They also employed BHT filtering; their subcategorisation lexicon listed not just valid SCFs for English verbs, but also their relative frequencies. This SCF acquisition system was further developed by Preiss, Briscoe and Korhonen (2007), improving accuracy and adding the ability to analyse the subcategorisation of nouns and adjectives.

Korhonen (2002) extends this work by exploring ways to use back-off models to improve the effectiveness of hypothesis testing. She compared the performance of backing off to an unconditional SCF distribution with backing off to the SCF distribution of the verb's semantic class, as indicated by the most frequent sense of the verb in WordNet (Fellbaum, 1998). She concludes that backing off to the unconditional prior is worse than not backing off at all. Korhonen's work resulted in the VALEX resource (Korhonen, Krymolowski and Briscoe, 2006)[1], containing subcategorisation frame frequency information for 6,397 verb types.

Sarkar and Zeman (2000) collected 137 SCFs on Czech verbs, using the manually-labelled Prague Dependency Treebank (PDT, Hajič

1 http://ilexir.co.uk/applications/valex/

and Hladká, 1998). They tried to learn the argument-adjunct distinction by generating all possible subsets of observed SCFs and using various kinds of hypothesis testing, including BHT, to find argument-containing frames that best explained the data.

Subcategorisation acquisition systems have also been described for Modern Greek (Maragoudakis, Kermanidis and Kokkinakis, 2000), Dutch (Spranger and Heid, 2002), Spanish (Esteve Ferrer, 2004), French (Chesley and Salmon-Alt, 2006; Messiant and Poibeau, 2008; Rambelli et al., 2016), Italian (Lenci et al., 2008; Lenci, Lapesa and Bonansinga, 2012), Urdu (Raza, 2011), Persian (Aminian, Rasooli and Sameti, 2013), and Brazilian Portuguese (Scarton et al., 2014).

In German, early work was performed by Wauschkuhn (1999), who semi-automatically collected subcategorisation observations on 1,044 German verbs, and later performed a manual evaluation of seven verbs. Eckle-Kohler (1999) also used semi-automatic techniques to obtain statistics on 6,305 verbs; she also made use of a manual evaluation.

The first automatic system for subcategorisation acquisition in German was developed in a concerted research project (Schulte im Walde, 2002a,b; Schulte im Walde and Brew, 2002; Brew and Schulte im Walde, 2002; Schulte im Walde, 2003, 2006). The work presented in chapter 5 builds on this project in several particulars, and specifics of these articles will be explored in further detail; I give a brief summary here. Schulte im Walde (2002a) presents the SCF acquisition system; this used a manually written grammar to train a head-lexicalised PCFG parser like the English parser described by Charniak (1997) on 18.7M words of newspaper text. The grammar rules circumscribe an inventory of 38 SCFs, and the trained parser model transparently encoded the relative frequencies of SCFs for the verbs in the corpus. Schulte im Walde collected SCF occurrence numbers for 14,229 verbs; Schulte im Walde (2002b) then presented a large-scale automatic evaluation against a dictionary, the *Duden Stilwörterbuch*. Schulte im Walde and Brew (2002) used an unsupervised clustering algorithm to group verbs together based on SCF frequency information extracted from the lexicon; the automatically-produced clustering was evaluated against a small manually-produced gold standard clustering of 57 German verbs in 14 classes. Brew and Schulte im Walde (2002) repeated this study using spectral clustering (Ng, Jordan and Weiss, 2002). Schulte im Walde (2006) repeated these experiments again after manually expanding the gold standard used for evaluation to 168 verbs in 43 classes, and explored three different levels of SCF information, with the third and most detailed level including selectional preference information. The results of adding selectional preferences were inconclusive, and I will pick up this thread again below in section 4.3. The section immediately following this dives deeper into the business of automatic verb classification.

Schulte im Walde (2009) gives a relatively recent survey of the literature on SCF acquisition.

4.2 AUTOMATIC VERB CLASSIFICATION

As we have seen in section 2.3.5 and section 3.5, verb classes cluster verbs according to their syntactic behaviour and semantic properties, such that two verbs belonging to the same class tend to license the same constructions, and can also be said to share one or more facets of their meaning. Verb classes can thus permit an efficient representation of the semantic space of verbs, since common meaning components need only be specified once per class. Furthermore, verb classes can have predictive value; the knowledge that an uncommon verb belongs to a particular class can be used to guess that it will enter into particular syntactic constructions, even if these have not been observed in real data.

Manual definitions of verb classes have been constructed for several languages including English (Levin, 1993; Kipper-Schuler, 2005), Spanish (Vázquez et al., 2000), Spanish and Catalan (Aparicio, Taulé and Martí, 2008), Czech (Pala and Horák, 2008), Mandarin (Liu and Chiang, 2008), and Arabic (Mousser, 2010, 2011). Verb class information has been successfully used for applications such as WSD (Dorr and Jones, 1996), MT (Dorr, 1997; Habash, Dorr and Traum, 2003), SRL (Swier and Stevenson, 2004; Shi and Mihalcea, 2005; Zapirain, Agirre and Màrquez, 2008), document classification (Klavans and Kan, 1998), discourse parsing (Subba and Di Eugenio, 2009), and subcategorisation acquisition (Korhonen, 2002).

Because verb classes are so useful for NLP tasks, but require considerable time and effort to develop, research has focused on automatic verb classification. This enterprise aims to induce such VerbNet-style classifications automatically from data, allowing faster and cheaper development of resources, and easier adaptation to different languages and linguistic domains, such as biomedical texts. Automatically created verb classifications have proved useful for NLP applications (Shutova, Sun and Korhonen, 2010; Guo, Korhonen and Poibeau, 2011).

As I have already covered in the last section, the German subcategorisation acquisition system described by Schulte im Walde was evaluated using automatic verb clustering, judging against a manually-produced gold standard clustering. Schulte im Walde (2000) had previously experimented with automatically clustering English verbs to 30 Levin classes using subcategorisation information.

Like the work described in this thesis, an array of research projects have followed this modus operandi of classifying verbs based on their subcategorisation behaviour, often making additional efforts to capture selectional preferences in some way as well. Korhonen, Krymolowski and Marx (2003) performed an automatic verb clustering

of 110 English verbs into 34 Levin-style classes using SCF information derived from Briscoe and Carroll (1997)'s SCF acquisition system, paying special attention to the behaviour of polysemous verbs; the evaluation measures used were Schulte im Walde and Brew's APP (cf. section 5.5.3) and a new measure, modified purity. They found that polysemous verbs with a strong predominant sense behaved much like monosemous verbs; verbs with regular polysemy could also be correctly clustered; but homonymous verbs with irregular polysemy frequently ended up by themselves in singleton clusters. Kawahara, Peterson and Palmer (2014) induce 'polysemy-aware' verb classes. Starting by automatically parsing the English Gigaword corpus, they next employ a two-step clustering, first clustering verb instances into clusters representing semantic frames, and then later clustering these semantic frames into verb classes. They evaluate using the test set produced by Korhonen, Krymolowski and Marx (2003). Peterson et al. (2016) repeats this study, modifying the second stage of the procedure by using SemLink to relate verb types to probability distributions over VerbNet classes, thus encouraging the final clustering to better resemble VerbNet, and producing slightly better results on the evaluation. Esteve Ferrer (2004) performed automatic verb classification of 514 Spanish verbs with Ward's hierarchical clustering based on frequency information for 11 SCFs, using as a gold standard the manually written verb classes of Vázquez et al. (2000).

Over the years, the research community has begun to re-use particular test sets (usually derived from VerbNet), leading verb classification to be increasingly approached as a supervised labelling problem. This style of work has tended to focus on feature engineering: identifying the information sources about verbs that can help an automatic classifier score better on a particular shared task. Merlo and Stevenson (2001) classified 20 English verbs that could be optionally intransitive into three classes: unergative, unaccusative and object-drop, based on a set of five heuristic linguistic indicators that would capture a particular semantic property of a verb's subject (e. g., animacy, which was estimated as the percentage of instances where a verb's subject was a pronoun). They trained a decision tree that attained 69.8% accuracy over a baseline of 33.9%. Joanis, Stevenson and James (2008) classified 835 English verbs into 15 VerbNet classes using a battery of 224 morphosyntactic features, achieving an accuracy of 58.4% with a support vector machine (SVM). Among the features they tried were grammatical aspect, tense, and voice, but these were not found to be helpful. Sun, Korhonen and Krymolowski (2008) classified 204 English verbs into 17 Levin classes with SCF counts taken from the VALEX lexicon. A parametric model of the subcategorisation preferences of each verb class was able to obtain 64% accuracy. Ó Séaghdha and Copestake (2008) later repeated this study using a SVM with a radial basis function kernel using the Jensen-Shannon divergence (cf.

section 5.5.2), improving accuracy to 67%. Vlachos, Korhonen and Ghahramani (2009) used a Dirichlet process mixture model (DPMM) to cluster verbs from the test set of Sun, Korhonen and Krymolowski (2008). DPMM chooses the number of classes itself and also produces soft clusterings; because of this, the evaluation measure used is the V-Measure (Rosenberg and Hirschberg, 2007), making it difficult to compare to other work. Li and Brew (2008) explored features for classifying English verbs using the test set of Joanis, Stevenson and James (2008) and a Bayesian Multinomial Regression classifier, obtaining 66.3% accuracy; the best performing feature set was SCFs combined with collocates – words neighbouring the verb in a four-word window. Since 2008, research has achieved increasing consensus over a set of standard evaluation measures, whereby precision is defined as modified purity and recall as weighted class accuracy; the *F*-score is then reported, which is the harmonic mean of these two.

In languages other than German and English, a paradigm has developed whereby fragments of the VerbNet hierarchy are first manually translated to act as a gold standard for evaluation, and then a pipeline to enable SCF tagging is constructed. Sun et al. (2010) used the information in an existing valence lexicon to cluster 171 French verbs into 16 classes, which were manually translated from the classification of Levin (1993). Falk, Gardent and Lamirel (2012) later expanded on this, merging three valence lexicons as a source of SCF frequency data to cluster 2,183 French verbs, obtaining the best results when SCFs were parameterised for syntactic features (e. g., appearing in the middle voice, or taking a sentential complement) and an automatically guessed label representing which thematic role grid the verb would have in the English VerbNet. In Italian, Lenci (2010) used SCFs acquired from the 326M word corpus La Republicca (Baroni et al., 2004) parsed with a dependency parser (Attardi and Dell'Orletta, 2009), and selectional preferences represented with top-level classes from MultiWordNet (Pianta, Bentivogli and Girardi, 2002). Vulić, Mrkšić and Korhonen (2017) conducted verb clustering in six languages (French, Brazilian Portuguese, Italian, Polish, Croatian, and Finnish) using word embedding vectors. With the aid of a cross-lingual dictionary (PanLex: Kamholz, Pool and Colowick, 2014), they post-processed the monolingual word vectors to add information derived from the English VerbNet. They then clustered the modified verb vectors with spectral clustering or Ward's hierarchical clustering (these two methods performed equally well), beating the previous state of the art for verb classification in French and Brazilian Portuguese.

Working in a more psycholinguistic direction and using Bayesian techniques, Parisien and Stevenson (2010) develop a hierarchical Dirichlet process model to jointly learn SCFs and verb classes from syntactic features. Under their model, instances of verb-grammatical-relation pair are generated by verb subcategorisation frame slots, which are

generated by verb types, which are in turn drawn from verb classes that capture diathesis alternation behaviour. The model parameters are estimated from data derived from a corpus of child-directed speech parsed with an automatic dependency parser. They evaluate by measuring the model's ability to generalise to novel verbs, adding new instances to the trained model and running training further. The novel verb instances replicated the behaviour of the dative alternation as manually extracted from Levin (1993). The results suggest that the model is able to deduce from usage examples that novel verb instances exhibit the dative alternation, and that capturing verb classes is necessary for this effect. Parisien and Stevenson (2011) extend this work by adding an animacy feature and 14 semantic markers associated with the verb taken from VerbNet, modifying the first level of the hierarchical model to represent not just verb subcategorisation frames, but also collections of meaning components (e. g., Intransitive Motion, Transitive Change of State, etc.). With this change, they demonstrate that the model is able to infer semantic components of a novel verb on the basis of patterns of syntactic alternations

Some prior work has attempted to model diathesis alternations for verbs directly. McCarthy (2001) set out to identify alternations for English verbs using SCF-verb co-occurrence data; by manually linking the SCFs used by Briscoe and Carroll (1997)'s system to frame types defined by Levin (1993), she could identify which verbs might be taking part in a given alternation. Collecting information about the verb's argument preferences (in different grammatical relations to the verb under different alternations) yielded data that could provide empirical support for potential alternations of the verb.

For more review of prior work on automatic verb classification, please see Schulte im Walde (2009) and Sun (2012).

4.3 SELECTIONAL PREFERENCES

Research into selectional preferences is quite well developed, with a long history in NLP and several viable techniques and evaluation methods. Selectional preferences have been shown to be useful for several NLP applications, including WSD (Resnik, 1997; Stevenson and Wilks, 2001; McCarthy and Carroll, 2003; Ye and Baldwin, 2006) and SRL (Gildea and Jurafsky, 2002; Erk, 2007; Zapirain et al., 2013). An interesting study by Shutova, Teufel and Korhonen (2013) also used selectional preference violations for detecting metaphors in running text.

Early work focused on WordNet; as an ontology, its pre-constructed inventory of concepts neatly fits our intuitions about how selectional preferences should work. This started with Resnik (1997), who modelled selectional preference concepts as hyponymy hierarchies under particular synset nodes in WordNet. Since then, several other

WordNet-based models of selectional preferences have also been developed, all of which attempt to identify effective ways to segment the hierarchy in order to optimally represent particular concepts. The tree-cut model of Li and Abe (1998) uses the Minimum Description Length principle (MDL, Rissanen, 1978) to describe regions within WordNet's noun hierarchy and discover an optimal concept granularity; Clark and Weir (2001) also aim to discover good partitions of WordNet, using hypothesis testing to identify segment boundaries; and Ó Séaghdha and Korhonen (2012) use Bayesian modelling on top of WordNet to identify noun groupings.

Because selectional preference information offers a window onto verb semantics, modelling it should also provide useful features for automatic verb classification; research along these lines is especially relevant to my work. As I have already touched on, Schulte im Walde (2006) experimented with automatic verb classification on the basis of subcategorisation information; she also enriched this data by developing a model of selectional preferences built on GermaNet. Her model is simple compared to the other WordNet-based models mentioned above; in her system, verb arguments are just assigned to one of 15 top-level GermaNet synsets. Like Resnik (1997), she treats the issue of word senses. GermaNet has a sense inventory for every lemma, and each possible sense of a given lemma votes for a top-level synset to represent it; this procedure is described in detail in section 6.2.4. Under her verb clustering evaluation, her model combining selectional preferences and SCF was better than a model using only SCF, but the difference was not statistically significant. She also tried several different combinations of frame slots, such as considering only the subject of intransitive verbs and the object of transitive verbs, a combination which might be expected to capture behaviour such as the causative alternation; this produced slightly better performance, but the effect was still not statistically significant. Generally, the benefits of her selectional preferences model were not consistent, and some of the frame slot combinations she tried did not work at all, leading her to conclude that 'the 15 conceptual GermaNet top levels are not sufficient for all verbs' (Schulte im Walde, 2006, p 189).

Korhonen, Krymolowski and Collier (2008) approached verb clustering for the biomedical domain, using the Information Bottleneck (Tishby, Pereira and Bialek, 2000) and Pairwise Clustering (Puzicha, Hofmann and Buhmann, 2000) algorithms. They experimented with parameterising SCF information for various feature sets, including two models of selectional preference. The first of these was a simple lexical preferences model (I will implement a similar model in chapter 6), and the second used Pairwise Clustering to cluster argument nouns together into argument classes. They found that information about tense was helpful but voice (active vs. passive) was not; the best performance was obtained with SCFs with prepositions and selectional

preferences. The argument noun clusters performed exactly as well as the lexical preferences model.

Sun and Korhonen (2009) also added a selectional preference model to an SCF-based automatic verb classification in English. Their method represents noun types with numerical vectors and then automatically groups these vectors into clusters using spectral clustering in order to represent selectional preference concepts; I will re-implement a version of their algorithm in section 6.2.2, and the method is described in detail there. They were able to show a performance benefit from combining the selectional preferences (SP) model with SCF information over the baseline SCF-only model on two separate English verb clustering gold standards (Joanis, Stevenson and James, 2008; Sun, Korhonen and Krymolowski, 2008). Sun and Korhonen also implement a lexical preferences method and find that it performs very well, although not as well as the spectral clustering method. Qualitative examination of the induced noun clusters showed that these captured general semantic categories (human, building, idea, etc.); they also found good overlap between the induced clusters and the thematic roles and selectional restriction labels listed in VerbNet. Their best results were obtained with 10–16 nouns per cluster on average. Moreover, the best results were obtained by modelling only the subject relation, perhaps because of the large number of subjects compared to objects and indirect objects in regular English. Sun, McCarthy and Korhonen (2013) modified this procedure to add as new features to the verb vectors the frequency of pairs of SCFs, these being intended to capture diathesis alternations; This change is shown to deliver better verb clustering performance on the same test sets.

Scarton et al. (2014) adapted this technique to Brazilian Portuguese, using a gold standard of 540 verbs in 16 classes translated from the English VerbNet, with additional verbs added manually. SCF tags are collected from several corpora, and SP concepts are constructed using Sun and Korhonen's method. Using spectral clustering with the MNCut algorithm, Scarton et al. achieve their best F-score of 43%, low compared to the state of the art English 80% and French 55%.

Rooth et al. (1999) present a generative latent model of SP that assumes that both the verb and its argument are generated by a latent concept. Expectation-Maximisation (EM, Baum, 1972) is used to simultaneously induce both the concepts and verb-argument clusters, by estimating model parameters using corpus data.

Schulte im Walde et al. (2008) also simultaneously induce soft clusters for verbs as well as for their arguments. The verb clustering is represented by latent variables; a PCFG is used to probabilistically relate these latent variables to the verb's subcategorisation frame and the selectional preference class of the verb's argument; EM is then used to estimate these latent variables. Selectional preference concepts are represented with a tree-cut model on the WordNet hyponym hier-

archy similar to Li and Abe's. This article, while highly relevant to this dissertation, is difficult to compare to quantitatively, as the published results are given in terms of perplexity.

Reichart and Korhonen (2013) present an unsupervised method that takes grammatical relation count data from the BNC corpus parsed with the RASP parser and constructs a joint model of SCF and SP using a Determinantal Point Process. Hierarchical clustering is then used to cluster verbs into verb classes. Evaluation against a set of 277 verbs taken from VerbNet shows promising results, although the gold standard is specifically constructed for the task by the authors, making it difficult to compare this to other work.

Aside from automatic verb classification, selectional preference models have usually been evaluated using two paradigms: the pseudoword approach and the argument plausibility approach. The pseudoword disambiguation task (Yarowsky, 1993; Chambers and Jurafsky, 2010) comes from work on WSD; given two potential slot-filling argument words, one real and one contrived, the SP model must guess that the real word is more probable. Argument plausibility approaches regress the outputs of models against judgements solicited from human annotators about the acceptability of verb-relation-argument triples. Such plausibility lists were published, for example, by Trueswell, Tanenhaus and Garnsey (1994), McRae, Spivey-Knowlton and Tanenhaus (1998), Keller and Lapata (2003) and Padó, Padó and Erk (2007).

Erk (2007) and Erk, Padó and Padó (2010) present a memory-based learning model of SP. The principle here is to collect verb-argument frequency data on the basis of observations from some corpus (here, FrameNet is used); these data are then generalised to new and unseen words by making use of a word space model to define a semantic similarity score function. For an arbitary argument, the most similar corpus-derived observations can then be retrieved, allowing estimates of the new argument's plausibility. This approach can be applied in both directions along the grammatical relation, allowing search for verb types based on an argument, or argument types based on a verb.

Bergsma, Lin and Goebel (2008) present a discriminative model of SP. Here, a SVM learns to distinguish real verb-argument pairs seen in a corpus from fake verb-argument pairs that are invented at random.

Van de Cruys (2009) models the selectional preferences of binary predicates using tensors to record (verb, subject, object) count observations; the tensor is factorised using a relatively small number of dimensions (up to 300) to regularise the model and to improve its ability to generalise to unseen data.

Ó Séaghdha (2010), whose method I re-implement in section 6.2.5, uses latent Dirichlet allocation (LDA) to model the relation between a verb and one of its arguments (verb-object, noun-noun, and adjective-noun). Ritter, Mausam and Etzioni (2010) use LDA to model binary predicates.

Van de Cruys (2014) presents a neural model of selectional preferences for predicting the plausibility of a verb-object pairs and verb-subject-object triples.

Zhang et al. (2020) modify the BERT word embedding model (Devlin et al., 2018) to produce for each word a separate embedding depending on which grammatical relation the word appears in.

Metheniti, Van de Cruys and Hathout (2020) explore whether SP information can be found inside the BERT model. They first manually annotated plausibility scores on pairs of head and dependents words, with five types of syntactic relation. After, they checked whether these plausibility scores correlated with probabilities assigned by BERT.

Several studies have directly compared multiple models of selectional preferences against each other, as I will do in chapter 6. Brockmann and Lapata (2003) compared several GermaNet-based methods using syntactic plausibility data in German; and Ó Séaghdha (2010) compared three LDA-based SP models, including the method of Rooth et al., 1999. Ó Séaghdha and Korhonen (2012) used a plausibility evaluation to compare an LDA version of the method of Li and Abe (1998) and a selectional preference version of the WSD model reported by Boyd-Graber, Blei and Zhu (2007) with the methods of Resnik (1997) and Clark and Weir (2002).

For the curious reader, Light and Greiff (2002) offer a survey of early work in selectional preference modelling.

4.4 COMPUTATIONAL ASPECT

Lexical aspect is seen as a source of information that is valuable to natural language understanding tasks. In particular, a treatment of aspect is critical to applications that must deal with temporal progression (Costa and Branco, 2012a); examples include information extraction, question answering, and document summarisation. Aspect has also been used as an information source for computational semantic analysis (Caselli and Quochi, 2007), event annotation (Pustejovsky et al., 2010; Bittar et al., 2011; Caselli et al., 2011), discovering the rhetorical structure of text (Baiamonte, Caselli and Prodanof, 2016), and caption analysis (Alikhani and Stone, 2019).

An early and influential paper on computational aspect was presented by Siegel and McKeown (2000), which built on the work of Klavans and Chodorow (1992). Klavans and Chodorow first applied Dowty's aspectual tests (cf. sections 2.4.2 to 2.4.4) to a corpus of one million words. They automatically detected instances of verbs in the present progressive, and used the ratio of such progressive instances to total instances as a score for how dynamic (non stative) a given verb type was. Klavans and Chodorow also tried other tests for dynamicity, namely finding verb instances that are complement to the verbs 'force' and 'persuade'; and those modified by intentional adverbs like 'de-

Linguistic indicator	Example clause
frequency	(not applicable)
'not' or 'never'	They had **never** been camping in their lives.
temporal adverb	I rang **immediately** for an ambulance.
no subject	Passives are used to foreground the patient.
past/pres participle	. . . **asking** for more.
duration 'in'-PP	I found the solution **in fifteen minutes**.
perfect	She **has** not seen that before.
present tense	Sascha **bikes** to work.
progressive	He **is seeing** someone new.
manner adverb	They **happily** ate the food.
evaluation adverb	In my view, he was treated **unfairly**.
past tense	Christian **worked** at a bank.
duration 'for'-PP	She shouted **for five minutes**.
continuous adverb	The deal has been postponed **indefinitely**.

Table 4.1: Aspectual indicators of Siegel and McKeown (2000), adapted from Table 4, p. 602.

liberately' and 'carefully'. However, these did not work because the corpus used was too small.

Siegel and McKeown (2000) extend this idea to a set of 14 *aspectual indicators*, shown in table 4.1; these are syntactic constructions or adjuncts that may co-occur with a given verb instance. These indicators can be expected to reflect the aspectual structure of the verb, although in some cases the connection is indirect: For example, manner adverbs, while they may be used to detect dynamic verb phrases, are actually markers of *intentional events*, or 'agentive events' in Siegel and McKeown's terms. Siegel and McKeown also make more aspectual distinctions than did Klavans and Chodorow, taking the five-class inventory of Moens and Steedman (1988, cf. section 2.4.8).

Finally, Siegel and McKeown consider the *fundamental aspectual class* of a verb phrase. This is defined to be the aspectual class of a clause before any aspectual transform or coercion is applied, such that the fundamental aspectual class is a function only of the verb and its arguments.

Siegel and McKeown manually annotated clauses from a corpus for two aspectual distinctions, stativity and culmination, and trained several kinds of supervised classifiers (decision trees, logistic regression, genetic programming) to recognise these categories. I will compare directly against these experiments in sections 8.1.4 and 8.1.6, so it is worth going over them in detail here.

In their first experiment, Siegel and McKeown manually annotated 1,478 clauses from medical records as either states (16.2%) or events (83.8%); these labelled data were then divided into 739 test clauses and 739 training clauses. The clauses were annotated by one linguist, and so no inter-annotator agreement was measured. Using a decision tree, Siegel and McKeown obtained an accuracy of 93.9%, a significant improvement over the baseline accuracy of 83.8%. The classifiers were reported to attend strongly to manner adverbs and the *in*-PP durative indicator, both of which imply dynamicity.

In their second experiment, Siegel and McKeown manually annotated 615 clauses from novels as either non-culminated (36.7%) or culminated (63.3%); these data points were then apportioned into 306 test clauses and 307 training clauses. The clauses were manually annotated by two linguists; Siegel and McKeown report an inter-annotator agreement of 91% (i. e., 81 of 89 items), which I calculate to represent a Cohen's κ of around 0.8. With a decision tree, Siegel and McKeown were able to obtain 74.0% accuracy, a significant improvement over the baseline accuracy of 63.3%[2]. The perfect tense indicator was the most important feature to the classifiers, as on this task it strongly implied culmination.

Three more publications, by Zarcone and Lenci, Friedrich and Palmer, and Falk and Martin, tread ground very similar to Siegel and McKeown. In all three of these papers, the authors first manually constructed datasets of verbs in various languages by labelling these for aspectual features, and subsequently trained supervised classifiers on these datasets. This work is of compelling relevance to this dissertation, and I will make direct comparisons to these results in sections 8.1.3 to 8.1.5.

Zarcone and Lenci (2008) annotated 3,129 verb instances from the Italian Syntactic-Semantic Treebank (Montemagni et al., 2003) with Vendlerian classes (41% achievements, 26% accomplishments, 18% statives, and 13% activities). They trained a maximum entropy classifier using groups of features classed into three kinds: adverbials (time-span and durative, agentive, frequency, etc.); morphological (tense and grammatical constructions); and syntactic (e. g., passive voice, presence/absence of direct object, locative modifier, semantic features of subject and object). Using 10-fold cross-validation (cf. section 8.1) on their labelled dataset, they achieve an accuracy of 85.4% over a baseline of 79.8%.

Friedrich and Palmer (2014) classified English verbs at a token level as being either stative or dynamic, or possibly both; this annotation was accomplished following Siegel and McKeown's notion of fundamental aspectual class. Verbs annotated were chosen using the LCS database (Dorr et al., 2001); the authors selected verbs that had only stative senses, dynamic senses, or a mix of stative and dy-

2 (Siegel and McKeown, 2000, Table 16, p 618)

namic senses; the verbs 'have' and 'be' were excluded. Two annotators manually classified 6,161 clauses from the MASC corpus (Ide et al., 2010); this annotated corpus, called Asp-MASC, attests 4,163 different verb types. Inter-annotator agreement was measured at Cohen's $\kappa = 0.7$. A second annotated corpus was also created to more strongly represent verbs that can refer to both stative and dynamic situations; this consisted of 2,667 clauses from the Brown corpus (Kučera and Francis, 1967) with a main verb from a list of 20 highly frequent aspectually ambiguous verbs. Inter-annotator agreement on this corpus was measured at $\kappa = 0.6$.

Friedrich and Palmer conducted a series of experiments on training supervised classifiers with the Asp-MASC dataset; the first two of these are relevant to my purposes. In both cases, the classifiers employed made use of a variety of features. Linguistic indicator features were the features listed by Siegel and McKeown (2000). The affinity of these features for particular verbs was estimated across the English Gigaword Corpus (Graff et al., 2003). Distributional features were word vectors for verbs, representing the syntactic behaviour of verbs, also derived from the Gigaword corpus by Thater, Fürstenau and Pinkal (2011). Finally instance-based features were particular to a verb token, including part of speech, tense, progressive, perfect, voice, and WordNet-based features of grammatical dependants.

Friedrich and Palmer's Experiment One used ten-fold cross-validation on the Asp-MASC dataset; the random forest classifier (84.1% accuracy) attained a higher score than a baseline that memorised the majority label class for each verb (83.6%), but the difference was not statistically significant. The equivalent accuracy using a simpler baseline of always predicting the label of the largest class seen during training is 72.5%.

Experiment Two again used ten-fold cross-validation on the Asp-MASC dataset, but this time the folds were grouped for verb type, so that the verb types in each test fold would not have been seen in any of the corresponding training folds. Here, the logistic regression classifier attained 81.9% accuracy; the dataset is the same as that used in the first experiment, so the baseline accuracy is also 72.5%.

Falk and Martin (2016) annotated entries from a valency lexicon of French verbs, Les Verbes Français (Dubois and Dubois-Charlier, 1997); they chose 167 frequent verbs, balanced for Vendlerian aspectual class. The lexicon contains a semantic decompositional representation of the meaning of each reading of a particular verb, made up of semantic primitives that can be sorted into 14 semantic fields. One annotator classified 1,199 verb readings from the lexicon for aspectual class. Falk and Martin take a very fine-grained view of telicity, and annotate verb instances in context using an eight point scale. These are:

1. strictly stative (S-STA, e. g., 'know')

2. stative with a dynamic reading (STA-ACT, e. g., 'think')
3. strictly dynamic and atelic (S-ACT)
4. variable telicity, compatible with both 'for' and 'in' adverbials (ACT-ACC)
5. weak accomplishments, which imply completion but are still compatible with 'for' adverbials (W-ACC, e. g., 'Peter filled the truck for one hour')
6. strong accomplishments, which are incompatible with 'for' adverbials (S-ACC, e. g., ? 'They broke the law for five days')
7. accomplishments that share some properties of achievements (ACC-ACH)
8. strict achievements (S-ACH)

These eight classes are then further organised into 3 high-level groupings: atelic (classes 1–3, representing 35.4% of the annotated corpus), variable telicity (class 4, 16.3%), and telic (classes 5+; 48.4%). The authors note that it is often found that different instances of a given verb lemma have different readings belonging to different aspectual classes.

Falk and Martin next train and test supervised classifiers on their annotated corpus with 10-fold cross-validation, using the three-way coarse-grained aspectual labels as targets. The classifiers made use of 38 features, comprising morphosyntactic and semantic features from the verb lexicon; these cover most of the features used by Siegel and McKeown (2000) and Zarcone and Lenci (2008). The baseline method that always predicts the majority class (telic) achieved 48.4% accuracy. In contrast, the best performing classifier (K* memory based learning) achieved 67.5% accuracy.

They then performed a task-based evaluation, following Costa and Branco (2012a, discussed below), constructing three tasks to mirror the structure of the TempEval task, but using the French TimeBank (Bittar et al., 2011) as labelled data with 10-fold cross-validation. By adding features estimated from the verb lexicon representing a verb's preference to be telic or atelic, on top of a baseline set of features derived directly from the TimeBank, they were able to show improvements in accuracy of 1–3% with a series of supervised classifier methods (decision trees, SVM, etc.).

Ranging further afield, and with less immediate relevance to my present project, another collection of work takes as its focus a wider notion of 'situation type', more relevant to sentence understanding and discourse processing.

Palmer et al. (2007) manually annotated English sentences as coding for different ten different situation types, such as events, states,

reports, generalising sentences, questions, and imperatives. They do not report measures of inter-annotator agreement. They then train a maximum entropy classifier to automatically label sentences for situation types, achieving 50.6% accuracy with ten-fold cross-validation on their manually labelled data. Friedrich, Palmer and Pinkal (2016) also annotated English sentences for seven situation types and trained classifiers, reaching accuracies up to 76%.

Other studies are concerned with habituality and genericity. Mathew and Katz (2009) produce a corpus annotating sentences as either habitual or episodic; Friedrich and Pinkal (2015) use this corpus, and also annotated some new sentences as being habitual, episodic, or stative. Govindarajan, Van Durme and White (2019) also construct a large dataset that labels sentences as being episodic, habitual or generic. Louis and Nenkova (2011) annotate sentences as specific or general ('broad statements about a topic'), and Friedrich et al. (2015) annotate subjects as generic or specific, and also clauses as generic or specific.

Finally, a different research angle has investigated the impact that aspectual information can have on external NLP tasks, particularly on understanding temporal relations. Costa and Branco (2012a) defined a set of linguistic indicators for Portuguese and estimated the relative frequency of these indicators for different German verbs using Web searches. They used TimeBankPT (Costa and Branco, 2012b), a translation of the TimeBank corpus (Pustejovsky et al., 2003) to Portuguese, as labelled data. The 2007 TempEval competition (Verhagen et al., 2007) sets out to predict the type of temporal relations that hold between between events. Costa and Branco achieved good results on this evaluation, comparable to the best results reported on the English TempEval.

Part II

ARGUMENT STRUCTURE

5 SUBCATEGORISATION ACQUISITION

As discussed in section 2.3.5, Levin (1993) argues for a very strong link between argument structure and semantics. In particular, she used the presence and absence of regular alternations of SCFs with a particular verb as a key criterion for determining which verb class that verb belongs to.

To investigate the argument structure of German verbs, my supervisor and I created an automatic subcategorisation acquisition system for German. I use this system to ingest the SdeWaC corpus and assign to each finite verb a label representing that verb's SCF. As described in section 2.3.3, a SCF is the particular pattern that a verb is instantiated with, indicating the number and types of the verb's arguments. A SCF can combine various syntactic roles, such as subject, accusative (direct) object, embedded clauses, etc.

I have already given a review of previous work on subcategorisation acquisition in section 4.1; as I noted there, the work described in this chapter builds on the research conducted by Sabine Schulte im Walde. In particular, I will recreate the experiments described by Schulte im Walde and Brew (2002) and Schulte im Walde (2006) below in section 5.5.

At the same time as the work in this chapter was being done, Scheible et al. (2013) published another subcategorisation acquisition system for German; coincidentally, this work is startlingly similar to mine. It used the SdeWaC corpus, automatically parsed with the `mate-tools` dependency parser (Bohnet, 2010), and rule-based software to extract subcategorisation information from the parses. The SCF inventory used was different than the one I use here. Scheible et al.'s SCF lexicon was not directly evaluated in their paper, although it was used to improve the performance of a statistical machine translation (SMT) system (Weller, Fraser and Schulte im Walde, 2013).

This chapter continues in the following way: Section 5.1 presents the inventory of subcategorisation frame tags used in this work, and section 5.2 a description of the SCF tagger. In section 5.3, I explain how the tagger is used to construct a subcategorisation lexicon of German verbs by collecting counts of how often each verb in SdeWaC occurs with each SCF. This lexicon is evaluated intrinsically in section 5.4. In section 5.5, I present an extrinsic evaluation of the lexicon using automatic verb classification, the task of automatically inducing Levin-style semantic verb classes on the basis of corpus data. I will frame this experiment as a kind of standardised evaluation that can be used to measure the quality of some model of verb semantics, so that this

`k` copula construction (*Ich bin **Student***, 'I am **a student**')

`n` nominative noun phrase (***Das Glas** zerbricht*, '**The glass** breaks')

`a` accusative noun phrase (*Der Hund beißt **den Postträger***, 'The dog bites **the postman**')

`d` dative noun phrase (*Darf ich **Ihnen** helfen?*, 'May I help **you**?')

`r` reflexive pronoun (*Er wäscht **sich***, 'He washes **himself**')

`p` prepositional phrase (*Guck mal **aus dem Fenster**!*, 'Look **out the window**!')

`x` expletive *es* (***Es** regnet*, '**It** is raining')

`i` subordinated non-finite clauses (*Das scheint **zu funktionieren***, 'That seems **to work**')

`s-2` clausal complement with finite verb in first (e. g., questions or imperatives) or second (e. g., main clause) position (*Er sagt, **er macht das***, 'He says **he will do it**')

`s-dass` subordinated clausal complement with *dass*-complementiser (*Ich weiß, **dass er morgen Geburtstag hat***, 'I know **that it is his birthday tomorrow**')

`s-ob` subordinated *ob*-clause (*Ich frage mich **ob das wirklich wahr ist***, 'I wonder **if that's really true**')

`s-w` indirect *wh*-question (*Ich weiß nicht **wie sie das macht***, 'I don't know **how she does it**')

Figure 5.1: Inventory of SCF components, adapted from Schulte im Walde (2002a).

same evaluation setup will return later in this dissertation to examine other verbal semantic phenomena. Section 5.6 circles back to examine some topics unearthed during the development of the SCF tagger, and section 5.7 concludes.

5.1 SUBCATEGORISATION FRAME INVENTORY

This thesis uses the 38 different subcategorisation frames of Schulte im Walde (2002a, 2006), which were originally derived from a PCFG parser. Frames combine complements to the verb, which are classified into various categories, such as types of nominals (`n` for nominative subject, `a` for accusative object, `d` for dative object, etc.), prepositionals (`p`), and clausals (e. g., `i` for infinitival clause). Figure 5.1 shows

a list of these complement types, illustrated with examples. Complement types are combined to give a unitary tag, so that `n` represents an intransitive verb, `na` a transitive verb, and `nad` a ditransitive. I take Schulte im Walde's second level of detail, whereby SCFs containing PPs are further sub-specified for the preposition and case of the prepositional argument. For example, in:

(5.1) Wir benutzen Ihre Umfragedaten nicht für eigene Zwecke.
We use your survey data not for own purposes.
We will not use your survey responses for private purposes.

complements to the verb *benutzen* are the subject *wir*, the direct object *Umfragedaten* 'survey data', and PP headed by *für* 'for' and with its argument in the accusative case. Put together, this results in an SCF tag of `nap:für.Acc`.

Note that this inclusion of prepositional constituents in the subcategorisation frame would seem to violate the argument-adjunct distinction (cf. section 2.3.3), which considers only arguments to be obligatory complements to the verb. PPs are usually adjuncts, but here they are included in the SCF anyway. This choice is partly a pragmatic approach to a linguistic category that often seems to be more of a spectrum than a black-and-white difference; Schulte im Walde's PCFG grammar model did not allow for her to reliably distinguish between arguments and adjuncts, for example. Also, paying attention to prepositions seems intuitively promising. Schulte im Walde (2006, p 180), for instance, points out that verbs in the same verb class tend to agree on prepositional complements (e. g., *denken/glauben an* 'to think (of), to believe (in)') or patterns of modification (e. g., directional PPs for manner of motion verbs). Finally, even where SCF tags do capture completely optional adjunct information, this is likely still of net benefit: Sun, Korhonen and Krymolowski (2008), for instance, found that adjuncts are very informative for verb semantics. Joanis, Stevenson and James (2008) also consider the argument/adjunct distinction for subcategorisation acquisition, and report that capturing some adjunct information in subcategorisation frames does not seem to hurt performance on a verb classification task.

5.2 SUBCATEGORISATION FRAME TAGGER

The input to the tagger is a syntactic analysis of a German sentence, either manually annotated as in NEGRA and TIGER, or automatically predicted by a parser.

I wrote software to perform rule-based syntactic analysis to identify finite verb instances and collect their complements; this machinery is able to handle auxiliary and modal constructions, undo passive constructions, locate separable verb prefixes, and perform clause type classification. Additionally, some raising and control constructions

are able to be analysed (e. g., *anfangen* 'begin'). These low-level syntactic analyses work on the level of grammatical relationships, modelled here as paths taken through the parse tree from one syntactic node to reach the other. I wrote a tree search engine for the NLTK project (Bird, Klein and Loper, 2009) that uses TGrep2 expressions to describe these parse tree paths; TGrep2 (Rohde, 2001) is a tool for finding structures in phrase structure trees that match a specified pattern. I later extended this search engine to support dependency graph structures.

The SCF tagger operates using a set of 17 rules to classify the verb's complements; these rules are listed in figure 5.2. TIGER-style edge labels (cf. section 3.1.1) specify the syntactic role of a constituent (e. g., `SB` for grammatical subject; see table 5.1 for a complete list of labels) and are a key source of information for the tagging task.

The tagger's rules result in SCF tags listing a maximum of three complements, following the SCF inventory introduced in the previous section. The tagger tries to prioritise argument-ish complements. For instance, PPs are treated with low priority, and are included in a tag only if there is no subordinate clause, since PPs are often adjuncts. Nominal adjuncts as well as clausal adjuncts (relative clauses and parentheticals) are always ignored.

5.3 SUBCATEGORISATION LEXICON

Running the SCF tagger over all 880M words of SdeWaC parsed with the `mate-tools` parser results in SCF tags for 82,873,358 verb instances. The resulting SCF lexicon lists 331,073 verb lemmas in total, although these contain a considerable degree of noise due to spelling errors and the parser's automatic analyses. Fortunately, the Zipfian distribution of the lexicon means that filtering is quite effective for removing noise (this topic will be revisited shortly in section 5.4). For example, only 3,935 verb lemmas appear more than 1,000 times, and instances of these lemmas together account for 95.6% of all verb tokens.

The lexicon includes 743 different SCF tags, including 212 that occur with regularity ($p \geq 10^{-4}$). Across the whole corpus, the most common SCF is the intransitive `n`, representing about 26% of all verb instances; this is followed by `na` (transitive, 14%), `ni` (intransitive with infinitival clausal complement, 14%), and `k` (predicative, 4%).

Table 5.3 shows an extract of the SCF lexicon for the German verb *halten*, with 350,182 instances the 26th most common verb in SdeWaC. Much like its English cognate 'hold', *halten* may denote a variety of situations. The Duden dictionary lists twenty senses of the verb meaning, variously, to hold (fast); to support; to keep oneself or sth. on a path or in a position; to defend or control (a football goal, a fortress); to keep; to contain; to succeed; to maintain or uphold; to aim at; to keep animals; to have an opinion; to organise (a meeting); to remain;

1. Only look at finite verbs or at non-finite verbs that are governed by finite auxiliaries or modals.
2. Filter out those verbs that are not in a passive construction and do not have a subject.
3. Any complement that has an edge tag of `SB` (subject) or `SBP` (passive subject) is classified as a nominative argument (`n`), if the verb is not in a passive construction.
4. Filter out verb phrase (VP) relative clauses, modifiers, parentheticals, and conjuncts; it is assumed that these will be adjuncts.
5. Clausal complements are categorised according to the phrase type of the clause (i. e., one of `i`, `S-2`, `S-dass`, `S-ob`, or `S-wo`).
6. Prepositional complements are all of type `p`, unless the verb is in a passive construction and the complement is marked as being the passive's subject (*von/durch* 'by' prepositionals).
7. For nominal complements, filter out those with POS-tag `PRF` (reflexive personal pronoun) first; this becomes `r`.
8. (to 13) Categorise remaining NP arguments as nominative `n` (edge labels `SB` subject or `SBP` passive subject), expletive `x` (`EP`, expletive *es*), accusative `a` (`OA`, accusative object), dative `d` (`DA` dative or `OA2` second accusative object), or copular `k` (`PD` predicate). For passive verbs, expletives are are ignored, and subjects become accusative `a`.
14. Passive sentences get a dummy `n` if there is no subject.
15. Filter out `p` if there is a verbal complement that can be categorised.
16. If the SCF includes `k`, then discard all other complements from the SCF.
17. If the SCF includes one of the combinations `ad`, `ar`, `dr`, or `x`, then remove any `p` appearing in the SCF.

Figure 5.2: List of the 17 rules of the SCF classifier.

AC	adpositional case marker
ADC	adjective component
AG	genitive attribute
AMS	measure argument of adjective
APP	apposition
AVC	adverbial phrase component
CC	comparative complement
CD	coordinating conjunction
CJ	conjunct
CM	comparative conjunction
CP	complementiser
CVC	collocational verb construction (*Funktionsverbgefüge*)
DA	dative
DH	discourse-level head
DM	discourse marker
EP	expletive es
HD	head
JU	junctor
MNR	postnominal modifier
MO	modifier
NG	negation
NK	noun kernel element
NMC	numerical component
OA	accusative object
OA2	second accusative object
OC	clausal object
OG	genitive object
OP	prepositional object
PAR	parenthetical element
PD	predicate
PG	phrasal genitive
PH	placeholder
PM	morphological particle
PNC	proper noun component
RC	relative clause
RE	repeated element
RS	reported speech
SB	subject
SBP	passive subject
SP	subject or predicate
SVP	separable verb prefix
UC	unit component
VO	vocative

Table 5.1: List of edge labels in TIGER, adapted from Smith (2003).

Verb	Count	Frequency (%)
sein 'be'	13759577	16.6
werden 'become'	6931451	8.4
haben 'have'	5947614	7.2
können 'can'	3444427	4.2
sollen 'should'	1521336	1.8
geben 'give'	1272820	1.5
wollen 'want'	893436	1.1
müssen 'must'	848129	1.0
machen 'do'	803841	1.0
kommen 'come'	722876	0.9
gehen 'go'	609574	0.7
stehen 'stand'	596603	0.7
sagen 'say'	523393	0.6
finden 'find'	509522	0.6
lassen 'let'	453745	0.5

Table 5.2: Most common verbs in the SCF lexicon.

to stop. Many of these word senses are also common to the English verb.

The table captures the syntactic behaviour of the verb, and provides insights into the relative frequency of its various uses in a large corpus. Prototypically transitive (`na`), the verb nevertheless has common intransitive (`n`) senses (e. g., to remain, to stop). *Halten* is frequently used with prepositional modification showing judgement (`nap:für`, `nap:von`), location (`nap:in.Dat`, `nrp:in.Dat`, `nap:an.Dat`, `np:in.Dat`, `nap:auf.Dat`, `nap:bei.Dat`), path or goal (`nrp.an.Acc`, `nap:in.Acc`), or instrument (`nap.mit`). The dative can be used to introduce a benefactive (`nad`), which is also possible with a PP (`np:für`). The reflexive is employed for the sense of keeping oneself in particular state (`nr`, combined with an adjectival complement, which is not captured in the SCF). Near the bottom of the table are several uncommon and incorrect SCFs: `nar`, `nap:zu.Dat`, `nai`.

For comparison, I also constructed an SCF lexicon from the combination of the NEGRA and TIGER corpora (1.2M words total); this contains 5,334 verb types, 134,133 tokens, and 158 unique SCF types. As a manually-created resource, it is free of the noise from automatic processing that is built into the SdeWaC-derived SCF lexicon, but, as we shall see, this advantage is outweighed by the small size of the NEGRA/TIGER corpus.

SCF	Count	Frequency (%)
nap:für.Acc	57585	21.9
na	50701	19.3
nap:in.Dat	20885	8.0
n	12091	4.6
nrp:an.Acc	8897	3.4
nrp:in.Dat	5435	2.1
nap:von.Dat	5408	2.1
nap:mit.Dat	5048	1.9
nr	4581	1.7
nad	4004	1.5
nap:an.Dat	3942	1.5
np:in.Dat	3911	1.5
np:für.Acc	3354	1.3
nap:auf.Dat	3092	1.2
nar	2916	1.1
nap:in.Acc	2679	1.0
nap:zu.Dat	2650	1.0
nai	2599	1.0
nap:bei.Dat	2458	0.9

Table 5.3: Frequency of SCFs recorded for *halten* 'hold' in the SCF lexicon.

5.4 INTRINSIC EVALUATION

An *intrinsic* (sometimes *in-vitro*) *evaluation* attempts to directly measure the quality of an NLP component, usually using a pre-defined labelled data set to test if the system behaves in an expected way; by contrast, an *extrinsic* (*in-vivo*) *evaluation* measures the performance of some NLP component inside a larger system, using the performance benefit to some external application as a proxy for the correctness of the component.

For the case of subcategorisation acquisition, intrinsic evaluation involves checking whether the tagger's outputs match with human judgements. Such human annotations can be collected to create either a *token-based* evaluation, where a human annotates particular verb instances for their correct SCF tags, or a *type-based evaluation*, where the human assembles a canonical list of valid SCFs for a particular verb lemma.

During early work on the SCF tagger, development was driven by iteratively finding tagger errors on a token level, and correcting the system output. My supervisor and I annotated the SCF tags for several hundred verb instances taken from SdeWaC; however, we found it difficult both to achieve high inter-annotator agreement on these instances, and also to use the manual annotations to measure the precision and recall of the SCF tagger. Ultimately, we decided that a token-based approach is too low-level, and the type-based model is more appropriate for evaluating a subcategorisation lexicon.

A previous type-based evaluation was conducted by Schulte im Walde (2002b), who compared SCFs for 3,090 verbs in her lexicon automatically against a machine-readable version of a large German dictionary, the Duden Stilwörterbuch (Dudenredaktion, 2001), achieving $F_1 = 62.3\%$ for simple SCF tags (10% over a simple baseline score), and $F_1 = 57.2\%$ for tags parameterised for prepositions, including case information (8% above baseline). The lack of access to the dictionary, however, prevents me from repeating such an evaluation.

Going back further in time, Eckle-Kohler (1999) manually evaluated her semi-automatically acquired SCF lexicon against the Duden Gesamtwörterbuch. She examined 15 verbs with a range of corpus frequencies, including two that prominently feature clausal complements. Her SCFs include adjectival and adverbial complements, and include a more fine-grained analysis of clausal complements than is found here (e. g., she distinguished perfect infinitives from present infinitives). The evaluation is presented in a case-based manner, and Eckle-Kohler does not quote any figures; however, from her published tables, I calculate $P = 54.7\%, R = 61.8\%, F_1 = 58.0\%$.

For the purposes of a type-based evaluation, I manually converted the SCFs for each verb listed as valid in Eckle-Kohler's thesis to the subcategorisation frame tag inventory used here.

SCF	Count	Squared Count	Normalised
n	31673	1003178929	0.323825
na	26081	680218561	0.219574
nS-dass	25296	639887616	0.206555
nS-w	23931	572692761	0.184864
nS-2	8602	73994404	0.023885
ni	7810	60996100	0.019689
nS-ob	4394	19307236	0.006232
...	...	...	...

Table 5.4: Computing the filtered list of SCFs for the verb *wissen* 'know' according to the procedure outlined by Schulte im Walde (2002a).

As can be seen in table 5.3, the SdeWaC-derived subcategorisation lexicon contains some degree of noise, due to errors in the output of the statistical parser. For a meaningful evaluation of an automatically-acquired verb subcategorisation lexicon, it is therefore important to apply some kind of *filtering* or hypothesis testing (cf. section 4.1) to the raw system output prior to scoring. From the lexicon, I constructed a list of valid SCFs for each of the 15 verbs in Eckle-Kohler, using the filtering procedure described by Schulte im Walde (2002a): For each verb, the co-occurrence counts of that verb with all SCFs are squared and these squared counts are then L1-normalised. All SCFs above a given threshold are treated as attested in the SdeWaC corpus; SCFs below this threshold are rejected as noise. This procedure is illustrated in table 5.4, where, with a threshold of 0.01, the SCFs for the verb *wissen* 'know' would be `n`, `na`, `nS-dass`, `nS-w`, `nS-2`, and `ni`, but not `nS-ob` or any lower-ranked SCFs. SCFs common to both the filtered subcategorisation lexicon and the gold standard are treated as true positives; SCFs found only in the filtered lexicon but not in the gold standard are false positives; and false negatives are those present in the gold standard but missing from the lexicon. With a threshold value of 0.04, I obtain $P = 71.1\%, R = 40.3\%, F_1 = 51.4\%$. For reference, the baseline method of Schulte im Walde (2002a), where each verb is assumed to occur with the two frames `n` and `na`, gives $F_1 = 35.1\%$. The NEGRA/TIGER-based SCF lexicon scores $F_1 = 38.7\%$ with a threshold of 0.01.

I explored several other methods of filtering automatically acquired SCFs that were listed and described in Korhonen (2002, section 3.3). Binomial Hypothesis Testing (BHT, Brent, 1993) estimates the probability, modelled as a Bernoulli trial, for each SCF, that that SCF might be erroneously assigned to a verb instance; for example, BHT estimates that any single verb instance in SdeWaC is tagged erroneously as `np.in.Dat` in 0.63% of cases. Using these probability estimates, and

the observed counts of verb-SCF co-occurrences, BHT can calculate the likelihood that a verb-SCF observation is due to chance. Filtering is then accomplished by retaining only highly significant combinations (specifically, verb-SCF pairs where $p < 0.01$). The Log-Likelihood Ratio (LLR, Dunning, 1993) is a non-parametric test that estimates the strength of the association between a verb and an SCF. Finally, the *t*-test (Sarkar and Zeman, 2000) is a parametric version of the LLR using the normal distribution. None of these hypothesis testing methods worked better than Schulte im Walde's simpler filtering algorithm.

This type-based evaluation is not a standard or shared task, and so we must be cautious with our interpretations. One issue is that a number of SCFs that are actually grammatical are not represented by the examples listed in the Duden and so are missing from Eckle-Kohler's gold standard. An example of this is the top-ranked intransitive use (n) of *wissen* 'know' shown in table 5.4, which is marked as incorrect in this task. Nevertheless, the evaluation has delivered encouraging results: The SdeWaC-derived SCF lexicon scores $F_1 = 51.4\%$, which is of the same magnitude as results obtained by Eckle-Kohler and Schulte im Walde, albeit slightly lower. This score is 16.3% above the baseline, which compares well to Schulte im Walde's larger-scale evaluation. As expected, the NEGRA/TIGER-derived lexicon performs best with a lower threshold value (0.01), an indication that it contains less noise than the automatically-acquired lexicon; however, the NEGRA/TIGER lexicon is based on a smaller corpus than SdeWaC, and is thus overall of lower quality when measured against the Duden (only 3.6% above the baseline).

5.5 AUTOMATIC VERB CLASSIFICATION

As I have outlined in the previous section, the other way to assess the quality of a resource like the subcategorisation lexicon is with an extrinsic evaluation. Fortunately, there is a prior extrinsic task, used by Schulte im Walde and Brew (2002) and Schulte im Walde (2006) to evaluate her subcategorisation lexica. This task, automatic verb classification, can be duplicated to obtain an evaluation that can be meaningfully compared with previous work.

As covered in section 2.3.5, Levin (1993) showed that verb meanings can be grouped based on their syntactic behaviour; for example, only transitive verbs coding for a change of state can enter into the middle construction. By counting how often which SCFs occur with which verbs, we should be able to approximately model the diathesis alternations a verb permits. This, in turn, should shed light on any meaning components a particular verb might have. Section 4.2 provides a summary of related work on automatic verb classification.

For this study, I collected SCF tags for all finite verbs in the first three million sentences of the automatically-parsed SdeWaC (80 mil-

lion words, about 10% of the corpus). These instances co-occur with 673 different SCF tags.

The method used follows Schulte im Walde (2006):

1. Verbs are represented by numerical descriptions of their subcategorisation behaviour, which I term their *subcategorisation preferences* (cf. figure 5.3);
2. the verbs are automatically clustered using the *k*-means clustering algorithm (cf. section 3.6); and
3. the resulting clustering is then evaluated against a gold standard clustering, using some evaluation metric.

Section 5.5.1 introduces the gold standard used for evaluation in this task, and section 5.5.2 details the method used for clustering verbs. Section 5.5.3 introduces the evaluation measures used to judge automatic clusterings against the gold standard. Sections 5.5.4 and 5.5.5 present two clustering experiments and compare the performance of the SdeWaC-derived subcategorisation lexicon against previous work. Discussion of the results follows in section 5.5.6.

5.5.1 *Gold standard classification*

Schulte im Walde (2006, pp 162ff.) presents a small manually constructed classification of 168 German verbs to be used for the purpose of evaluation for automatic clustering experiments. The verbs are grouped into 43 classes based on shared meanings, and not always on the basis of identical syntactic behaviour; examples are Aspect (e. g., *anfangen* 'begin'), Propositional Attitude (e. g., *denken* 'think'), and Weather (e. g., *regnen* 'rain'). The classes were deliberately constructed to resemble the classification of Levin (1993). Examples of such similarities are Schulte im Walde's Aspect class, which resembles Levin's Begin; Schulte im Walde's Position, which is a sub-class of Levin's Dangle; and Schulte im Walde's Obtaining, which subsumes Levin's Get and Obtain classes. The classes also agree well with the German verb classification of Schumacher (1986) and the English FrameNet project (Baker, Fillmore and Lowe, 1998). Classes contain both high- and low-frequency verbs, and class size ranges between two and seven. Eight verbs included in the classification are semantically ambiguous, and belong to more than one class, reflecting their different word senses. Schulte im Walde (2006, p 189) notes that the manual classification is a difficult gold standard; for example, the Perception and Observation classes are syntactically and semantically very similar to each other, and should be difficult for an algorithm to distinguish.

Some of the 43 classes are further grouped into higher-level classes, which number 26. These coarser-grained classes are, however, not

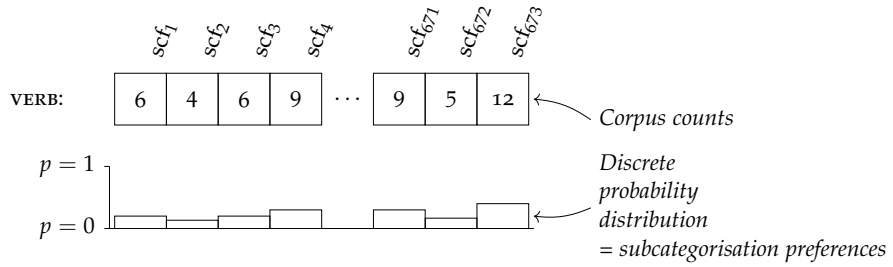


Figure 5.3: Subcategorisation preferences: Verbs as vectors of discrete probabilities over possible subcategorisation frames.

used by Schulte im Walde (2006) for her verb clustering experiments. Similarly, I follow her example and also disregard class-subclass relations, taking the sub-classes to be separate entities. An advantage of this choice is that this gives a set of verb classes of roughly similar size; by contrast, some of the coarse-grained top-level classes are disproportionately large. An example is the Transfer of Possession (Obtaining) class, which makes up 25% of the gold standard, even though its sub-classes 'Transfer of Possession (Giving)', 'Manner of Motion', and 'Emotion' are semantically very different to each other.

5.5.2 *Verb clustering*

The counts of SCFs for each verb lemma are normalised into vectors of conditional probabilities in the manner shown in figure 5.3. Each of these vectors expresses the subcategorisation preferences of a given verb v, and the elements of a vector are indicated with

$$P(\text{scf} = f \mid \text{lemma} = v)$$

for a given SCF f, or, more simply, the probability of an instance of v occurring with f.

These conditional probabilities are smoothed using backing off to the prior probability $P(\text{scf} = f)$; *Backing off* is a technique to smooth or combine different probabilistic models, interpolating between a noisy higher-level model with sparse data counts and a more reliable lower-level model with good coverage. Katz backing off (Katz, 1987) is typically used in n-gram language models; I use it in this chapter as if verb-SCF pairs were 2-grams of the form $\langle \text{scf}, \text{verb} \rangle$.

With verbs represented as discrete probability distributions, similarity between verbs is computed with two distance measures: the Jensen-Shannon divergence and the skew divergence. Both of these are variants of the Kullback-Leibler divergence (Kullback and Leibler, 1951):

$$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{5.1}$$

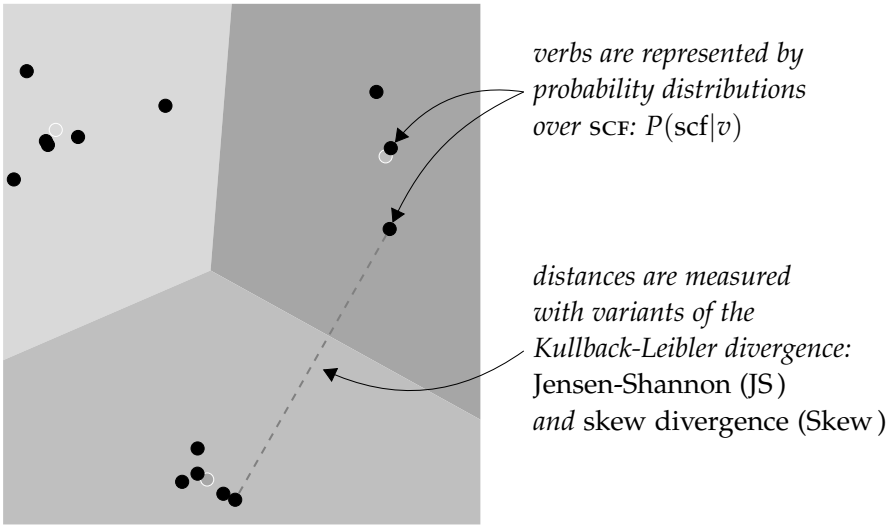


Figure 5.4: Illustration of the automatic verb clustering process.

This information theoretic measure calculates the dissimilarity between two discrete probability distributions p and q, where q is taken as the reference probability distribution. The Kullback-Leibler divergence is asymmetric, and $D = 0$ if p and q are identical. Furthermore, D is undefined if some q_i is zero but the corresponding p_i is not. The two variants used here both avoid this issue with zero probabilities:

$$\text{JS}(p,q) = D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2}) \tag{5.2}$$

$$\text{Skew}(p,q) = D(p\|\alpha q + (1-\alpha)p) \tag{5.3}$$

The *Jensen-Shannon divergence* (equation (5.2), Lin, 1991), also known as the *information radius*, solves this problem associated with zero probabilities in q by summing the divergences of p and q from their middle point, the mean of p and q. This renders the Jensen-Shannon divergence symmetric and always non-negative. The square root of the Jensen-Shannon divergence is a mathematical metric, called the Jensen-Shannon distance.

In contrast, the *skew divergence* (equation (5.3), Lee, 1999) is asymmetric. It uses an interpolation parameter α to ensure that there are no zero probabilities. In this work, I follow Schulte im Walde (2006) in setting $\alpha = 0.9$, even though the original author recommends values of 0.99 or greater for α.

5.5.3 *Evaluation measures*

This section introduces several measures of cluster purity for comparing clusterings to a ground truth, chosen to match Schulte im Walde and Brew (2002) and Schulte im Walde (2006) for ease of comparison.

I will take a hard clustering, such as one found by hierarchical agglomerative clustering or k-means, to be an equivalence relation that partitions n samples into k disjoint sets: $\mathcal{C} = \{C_1, \ldots, C_k\}$. The *adjusted Rand index* (Hubert and Arabie, 1985) models cluster similarity as a function of the overlap of a cluster C_i in the clustering $\mathcal{C}$ with another cluster G_j in the gold standard clustering $\mathcal{G}$. Letting this value be $\mathcal{CG}_{ij} = |C_i \cap G_j|$,

$$\text{Rand}_a(\mathcal{C}, \mathcal{G}) = \frac{\sum_{i,j} \binom{\mathcal{CG}_{ij}}{2} - \left[\sum_i \binom{|C_i|}{2} \sum_j \binom{|G_j|}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{|C_i|}{2} + \sum_j \binom{|G_j|}{2}\right] - \left[\sum_i \binom{|C_i|}{2} \sum_j \binom{|G_j|}{2}\right] / \binom{n}{2}} \quad (5.4)$$

This calculation adjusts for chance, and Rand_a ranges between zero for uncorrelated clusterings and one for identical clusterings.

Under the *pairwise F-score* (Hatzivassiloglou and McKeown, 1993), the gold standard clustering $\mathcal{G}$ is regarded as consisting of $\binom{n}{2}$ binary judgements of whether two verbs belong in the same cluster or not; a clustering under consideration $\mathcal{C}$ gives a similar set of judgements. With respect to the gold standard, these judgements in $\mathcal{C}$ can be viewed as correct or incorrect, and the agreements and disagreements listed in a confusion matrix. This permits the use of the common information retrieval metrics precision (P) and recall (R), as well as the F-score derived from these two:

$$\text{PairF}(\mathcal{C}, \mathcal{G}) = \frac{2P(\mathcal{C}, \mathcal{G})R(\mathcal{C}, \mathcal{G})}{P(\mathcal{C}, \mathcal{G}) + R(\mathcal{C}, \mathcal{G})} \quad (5.5)$$

The PairF is known to have a nonlinear response, penalising the first few mistakes more than later ones (Schulte im Walde, 2006). An advantage, however, is its use of a confusion matrix, which allows the statistical significance of differences in scores to be computed using McNemar's test (McNemar, 1947).

The *Mutual Information* measure (Strehl, Ghosh and Mooney, 2000) measures cluster similarity from the viewpoint of entropy; it is scaled according to the number of clusters, which should prevent the bias towards smaller clusters that is introduced because smaller clusters have higher purity.

$$\text{MI}(\mathcal{C}, \mathcal{G}) = \frac{1}{n} \sum_i \sum_j \mathcal{CG}_{ij} \frac{\log\left(\frac{n\mathcal{CG}_{ij}}{\sum_k \mathcal{CG}_{ik} \sum_k \mathcal{CG}_{kj}}\right)}{\log(|\mathcal{C}||\mathcal{G}|)} \quad (5.6)$$

Finally, the *Adjusted Pairwise Precision* (Schulte im Walde and Brew, 2002), like PairF, views equivalence relations as sets of pairs of items, and computes the information retrieval measure of precision, taking the precision of the whole clustering to be a weighted average of the

Data Set	Eval	Distance	Oracle	Random Best	Random Mean	Ward
Schulte im Walde	PairF	JS	40.23	1.34 → 16.15	13.37	17.86 → 17.49
		Skew	47.28	2.41 → 18.01	14.07	15.86 → 15.23
	Rand_a	JS	0.358	0.001 → 0.118	0.093	0.145 → 0.142
		Skew	0.429	−0.002 → 0.142	0.102	0.158 → 0.158
NEGRA/TIGER	PairF	JS	29.43	2.01 → 17.19	13.38	17.13 → 16.30
		Skew	35.08	2.76 → 14.21	13.43	16.67 → 17.49
	Rand_a	JS	0.268	−0.002 → 0.144	0.103	0.141 → 0.133
		Skew	0.324	−0.005 → 0.115	0.106	0.136 → 0.144
SdeWaC	PairF	JS	43.33	2.30 → 23.42	19.25	26.81 → 25.70
		Skew	59.42	2.64 → 22.45	19.49	26.16 → 26.13
	Rand_a	JS	0.393	0.001 → 0.196	0.154	0.229 → 0.219
		Skew	0.552	0.004 → 0.186	0.159	0.222 → 0.222

Table 5.5: Evaluation of the NEGRA/TIGER and SdeWaC SCF lexica on the clustering task of Schulte im Walde (2006).

precisions of each of its component clusters. Writing the precision of a given cluster C_i as $P(C_i, \mathcal{G})$,

$$\text{APP}(\mathcal{C}, \mathcal{G}) = \frac{1}{|\mathcal{C}|} \sum_i \frac{P(C_i, \mathcal{G})}{|C_i| + 1} \tag{5.7}$$

5.5.4 *Experiment 1*

This experiment duplicates the setup of Schulte im Walde (2006), clustering the 168 verbs in the gold standard using k-means clustering into $k = 43$ classes (i. e., the k-means algorithm is set to match the number of clusters in the gold standard)[1]. I compare to Schulte im Walde's second level of SCF granularity, with PP head and case information, which is the format delivered by the SCF tagger as described in section 5.1. The cluster purity measures used in this experiment are the PairF and the Rand_a.

Results of the clustering experiment are presented in table 5.5. For comparison, results published by Schulte im Walde (2006, p 174, Table 7) are reprinted in the first row under 'Schulte im Walde'. The second row of the table shows the performance of the NEGRA/TIGER-derived subcategorisation lexicon on the task; note that only 160 of the 168 verbs are attested in NEGRA/TIGER (17,312 tokens total), which lowers the maximum PairF or Rand_a scores achievable by this lexicon. The lexicon derived from the first 80M words of SdeWaC is scored in the third row; this corpus contains all 168 verbs (1,043,505 instances total).

The clustering task allows for several methods for initialising the k-means centroids:

1 Schulte im Walde (2006, p 181) notes that in her experiments, she achieves optimal results with $k = 71$.

RANDOM k verbs are taken at random and used as the initial cluster centroids (Forgy initialisation);

HIERARCHICAL CLUSTERING agglomerative hierarchical clustering is used to put the objects into k groups; the centroids of these groups are then used as the initial cluster centroids; and

ORACLE the k-means algorithm is initialised with centroids calculated from the gold standard clusters.

Random initialisation is repeated ten times; the 'Random Mean' value in table 5.5 records the average value of these trials, and the 'Random Best' value shows the trial with the least total intra-cluster divergence (not the trial with the best evaluation result). The Random Best column also shows both the starting and ending cluster purity scores.

Here I use only the best criterion for agglomerative hierarchical clustering as found by Schulte im Walde (2006): Ward's criterion[2]. Hierarchical clustering is a greedy bottom-up algorithm, and its output is deterministic, meaning that one trial is enough to measure performance under this configuration. As with the Random Best column, the 'Ward' column shows starting and ending cluster purity scores.

For ease of interpretation, a random baseline is computed as the average of the scores of 50 random partitions: PairF $= 2.08$, $\text{Rand}_a = -0.004$. Likewise, an optimal baseline can be computed by evaluating the gold standard clustering against itself. Note that a perfect score cannot be achieved here, because there are polysemous verbs included in the gold standard, the various senses of which are grouped into different clusters. Therefore, in calculating the optimal baseline, a sense is picked at random for each such polysemous verb; the average over 50 such trials then gives PairF $= 95.81$, $\text{Rand}_a = 0.909$.

The table shows that the performance of the NEGRA/TIGER-based SCF lexicon on the verb classification task is on par with the values reported by Schulte im Walde, with the exception of the 'Oracle' condition.

I also observe that the hierarchical clustering with Ward's criterion used for initialisation works better than the subsequent k-means clustering for SdeWaC, as well as for the variants of NEGRA/TIGER using the Jensen-Shannon distance measure.

Statistical significance can be computed on the 'Random Mean' scores by running many k-means trials initialised with random partitions; these scores are a function of the random cluster initialisation and can be taken to be another random variable. The scores are generally normally distributed, but I apply a Box-Cox transform (Box and Cox, 1964) to remove any kurtosis or skewness from the data, before

2 I did explore the other methods of hierarchical clustering used by Schulte im Walde (2006); my results mirror hers, and Ward's performs best on this task.

	JS		Skew	
Dataset	APP	MI	APP	MI
Schulte im Walde and Brew	0.144	0.357	0.114	0.320
NEGRA/TIGER	0.109	0.322	0.107	0.323
SdeWaC	0.159	0.367	0.158	0.371

Table 5.6: Evaluation of SCF lexica on the clustering task by Schulte im Walde and Brew (2002) using random k-means cluster initialisation.

comparing them to the reported 'Random Mean' value from Schulte im Walde (2006) using a standard t-test.

The SCF lexicon based on NEGRA/TIGER performs comparably to Schulte im Walde: The score for the PairF/JS condition is not significantly different; the PairF/Skew condition is significantly lower; and both Rand_a conditions are significantly higher at at least the $p < 0.01$ level. The lexicon based on SdeWaC performs significantly better than Schulte im Walde with all distance measures and cluster purity scores ($p < 0.001$).

5.5.5 *Experiment 2*

This experiment described in this section replicates the evaluation done by Schulte im Walde and Brew (2002), which is very similar to the experiment from the previous section. The differences are that the gold standard for this task lists only 57 verbs clustered into 14 classes; these 14 classes[3] are a proper subset of the 43 classes from the full manual verb classification. Again, the clustering algorithm is run using $k = 14$, matching the number of classes in the gold standard. Clusters are always initialised randomly.

Evaluation measures used are the APP and MI. The random baseline is calculated as the average of 10 random clusterings: $\text{APP} = 0.017, \text{MI} = 0.229$. This smaller gold standard does not include polysemous verbs, so the calculation of the optimal baseline is uncomplicated: $\text{APP} = 0.291, \text{MI} = 0.493$.

Results of the experiment are shown in table 5.6. As before, I reprint the results published by Schulte im Walde and Brew (2002, Fig. 2, p 228) in the first row of the table. The NEGRA/TIGER-derived lexicon is shown in the second row; only 54 of the 57 verbs from the gold standard are attested in NEGRA/TIGER (6,765 instances). The third row shows the evaluation of the SdeWaC-derived lexicon, which contains all 57 verbs (425,660 instances).

3 Aspect, Propositional Attitude, Transfer of Possession (Obtaining), Transfer of Possession (Supply), Manner of Motion, Emotion, Announcement, Description, Insistence, Position, Support, Opening, Consumption, and Weather.

I compute statistical significance with a *t*-test as in Experiment 1. Compared to the results of Schulte im Walde and Brew, the SCF lexicon based on NEGRA/TIGER scores significantly lower with the JS/APP, JS/MI, and Skew/APP configurations, and significantly higher with Skew/MI, all at the $p < 0.001$ level. The SCF lexicon based on SdeWaC scores higher under all conditions ($p < 0.001$).

5.5.6 *Discussion*

The two experiments outlined in this section have both demonstrated that the SCF lexicon based on SdeWaC works better on the automatic verb classification than the resource originally developed by Schulte im Walde. This result is not surprising, because the SCF lexicon I have described here is constructed using much more data (Schulte im Walde used a corpus of 35M words), which should give more accurate estimates of verbal subcategorisation preferences. This effect should be particularly true for low-frequency verbs; indeed, Schulte im Walde (2006, p 184) complains that low-frequency verbs with their associated noisy probability distributions 'destroy' the coherence of the clustering.

The better performance on the extrinsic evaluation leaves open the possibility that the SCF acquisition system I have described here functions better than Schulte im Walde's. This, too, is entirely conceivable: The `mate-tools` statistical parser is trained on TIGER, while Schulte im Walde used a hand-written PCFG for syntactic analysis. The quantity of manually-constructed linguistic annotation in the NEGRA/TIGER corpus surely must be greater than the syntactic intuitions captured in Schulte im Walde's PCFG. We can also expect the `mate-tools` parser to be more robust to different types of language and different domains.

We have also seen in both experiments that the SCF lexicon based on NEGRA/TIGER performs worse than Schulte im Walde's. This finding is also expected, since the NEGRA/TIGER corpus at 1.2M words contains many times fewer verb instances than Schulte im Walde's 35M words or the 80M words of SdeWaC used in this section. The smaller number of verb instances inevitably causes problems with data sparsity; we have also seen that NEGRA/TIGER does not attest all of the verbs included in Schulte im Walde's gold standard.

The PairF and Rand_a evaluation metrics used in Experiment 1 are almost perfectly correlated with each other. From the values in tables 5.5 and 5.8, I calculate Pearson's $r = 0.998$. Similarly, both evaluation measures used in the second experiment (APP and MI) are also highly correlated; tables 5.6 and 5.8 show that $r = 0.994$.

Skew divergence seems to help in experiment 1 and hurt in experiment 2, but a *t*-test shows that neither of these effects are statistically significant with $p = 0.140$ (Experiment 1) and $p = 0.197$ (Experiment

2). Skew divergence and Jensen-Shannon divergence are highly correlated with each other, with Pearson's $r = 0.989$ (Experiment 1) and $r = 0.986$ (Experiment 2).

5.6 DEVELOPMENT OF THE TAGGER

5.6.1 *Edge labeller*

At first, the SCF tagger was developed against the manually-annotated corpora NEGRA and TIGER. Later, to allow it to collect subcategorisation information on a larger corpus like SdeWaC, I initially relied on automatic parses from constituency parsers, beginning with the Stanford parser, and later switching to the Berkeley parser (cf. section 3.1.3). Both of these parsers are distributed with built-in models trained on the TIGER corpus.

Constituency parsers produce syntactic trees showing phrase structure, like in figure 3.1. These trees are annotated with the POS tags of the internal tree nodes (i. e., phrase structure categories); however, the parser does not predict the edge labels connecting these internal nodes. As discussed in section 3.1.1, edge labels are indispensable for distinguishing the grammatical roles of German verb-argument pairs. As the rules listed in figure 5.2 make clear, my SCF tagger makes use of the edge labels annotated in TIGER to:

- Identify subject constituents (marked as `SB/SBP`; can be verbal, clausal, adjectival, adverbial, prepositional, or nominal phrases);
- Identify expletive uses of *es* 'it' (a semantically null subject, marked as `EP`);
- Morphologically classify objects as accusative or dative objects (`OA/DA/OA2`);
- Identify copular constructions (nominal complement tagged with `PD`); and
- Filter out verbal adjuncts (relative clauses, parentheticals, and modifier clauses, tagged with `RC/MO/PAR`).

To supply grammatical role information to the SCF tagger, I created an automatic edge labeller. This tool was able to ingest the output of a constituency parser, and assign each verb-complement relation an edge label. Figure 5.5 shows the architecture of the subcategorisation acquisition system including the edge labeller.

Using the NEGRA and TIGER corpora as labelled datasets for modelling this classification task allowed the use of a supervised training paradigm. To accomplish this, every finite verb instance in NEGRA and TIGER was collected; further, for each of these, all of the verb's complements were collected. Each verb-complement pair is then a data

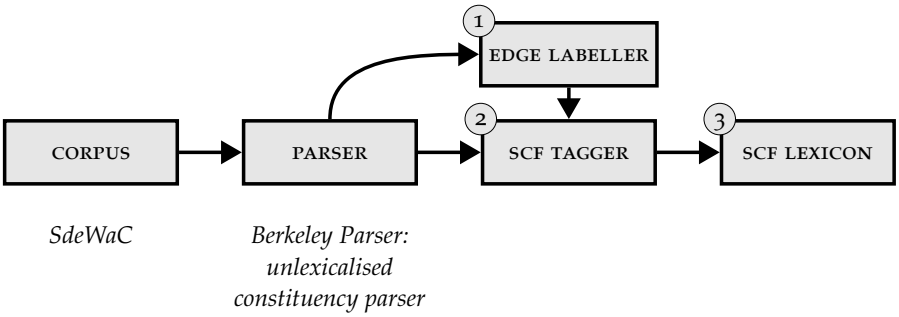


Figure 5.5: Early architecture of the SCF tagger system.

Category	POS	NP head	Prep.	Verb	Position	Clause type	Article
SB	n	mann	—	jagen	left	S-2	der
OA	n	wolf	—	jagen	right	S-2	den
MO	p	wald	in	jagen	right	S-2	den

Table 5.7: Some extracted features for the verb arguments in Figure 5.7.

point, and its manually annotated edge label as given in NEGRA or TIGER is the desired output of the edge labeller. In NEGRA/TIGER there are 337,275 of these labelled pairs; these labelled pairs were divided at random into a training set (90%) and a test set (10%). The data points, consisting of a verb-complement pair, are converted into numerical and categorical features, so that they can be learned by an automatic classifier.

Features for classification used by the edge labeller are listed in figure 5.6, and table 5.7 illustrates how the verb arguments in figure 5.7 are converted into feature vectors. The features selected reflect several intuitions about the edge labelling task:

1. Complements to a verb can be nominal, prepositional, or clausal.

2. *n*-grams on the end of the unlemmatised nominal head, or any modifying adjective thereof, can be useful for inferring case, as can the form of an article specifying the nominal head.

3. Prepositional complements can represent sentential subjects if the verb is in the passive form (when headed by *von* 'by' or *durch* 'through').

4. The leftmost daughter node of a constituent can indicate clausal relationships for verbal complements, helping, for example, to detect relative clauses.

Using a maximum entropy (Berger, Della Pietra and Della Pietra, 1996; Abney, 1997) multi-class classifier with L2 regularisation ($\lambda^{-1} = 3$), I was able to obtain an *F*-score of 95.5% on the test set. This is a

1. (category) The lemma of the verb;
2. (category) The POS of the complement;
3. (category) The clause type of the verb;
4. (left/right) The location of the complement relative to the main (finite) verb of the clause;
5. (category) The lemma of the complement's syntactic head;
6. (category) The form of definite or indefinite article specifying this syntactic head;
7. (categories) n-grams on the end of the unlemmatised syntactic head ($n \leq 4$);
8. (categories) The adjective modifying the syntactic head, and n-grams on the end of the unlemmatised form of this adjective ($n = 3$);
9. (yes/no) Whether the complement occurs to the left of a reflexive pronoun in the sentence (nominal complements to the left of a reflexive should almost always be subjects);
10. (yes/no) Whether the complement occurs to the right of a noun phrase in the sentence;
11. (yes/no) Whether the complement is sister to a pronoun or an NP with a definite, indefinite, or possessive article;
12. (cardinal) The distance of the complement from the left clause boundary;
13. (yes/no) Whether the verb is in the passive form or not;
14. (category) The lemmatised form of the preposition, if the complement is a PP;
15. (categories) The lemmatised form and POS of the main verb, as well as the clause type, if the complement is clausal; and,
16. (categories) The form and POS tag of the first word of the clause, if the complement is clausal.

Figure 5.6: Features to the edge labeller for classifying the complement of a verb.

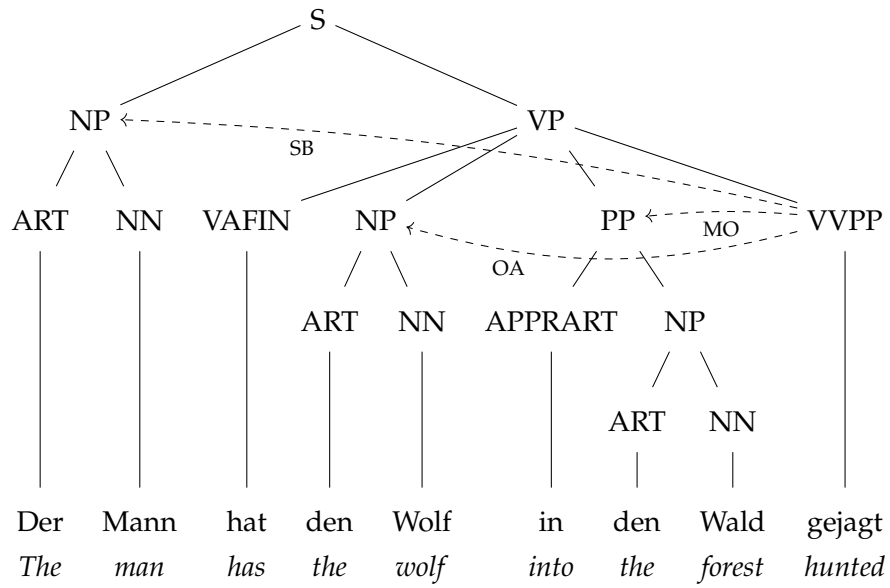


Figure 5.7: Syntactic roles of arguments to the verb indicated using TIGER edge label tags. 'The man chased the wolf into the forest'.

very good result, but it turns out that the edge labeller needs one more refinement before it can be used effectively for its intended purpose.

In a naïve setup, such as the evaluation performed using the NEGRA/TIGER-derived test set, the edge labeller makes the strong assumption that each complement to a verb can be assigned an edge label independently of the labels of the verb's other complements. When applied to actual automatic parser output, however, this leads the edge labeller to predict impossible situations, such as when two complements to the same verb are both labelled as being subjects. This type of error naturally arises because subjects have a very high prior probability in the edge labelling task: Subjects make up 40% of all verbal complements in NEGRA/TIGER, and are more than three times as common as the next most frequent complement type, accusative direct object.

To remedy this, the set of a verb's complements is collected, and the probability of each edge label for each complement is computed. Then, the combination of edge labels is chosen that maximises the joint probability of all the labellings, using Viterbi decoding (Viterbi, 1967). The search is constrained to guarantee that:

1. There is at most one subject of the verb; and
2. there is at most one accusative object of the verb,

under the assumption that the joint probability is the product of the individual label probabilities.

Date		19.7.13	20.7	22.7	17.10	5.11	30.1.14	3.2	6.2	8.5
SiW2006	PF/IR	15.90	16.45	17.35	17.65	16.57	18.26	18.82	19.21	19.25
	PF/SK	15.78	16.47	17.38	18.04	17.94	17.80	18.19	19.92	19.49
	RA/IR	0.124	0.131	0.136	0.142	0.129	0.146	0.151	0.153	0.154
	RA/SK	0.125	0.131	0.138	0.146	0.144	0.143	0.146	0.165	0.159
SiWBrew2002	APP	0.133	0.141	0.162	0.161	0.162	0.146	0.138	0.128	0.159
	MI	0.337	0.345	0.379	0.364	0.373	0.366	0.351	0.351	0.367
Duden	F_1				0.456	0.487	0.492	0.526	0.528	0.531

Table 5.8: Evaluation of the SCF lexicon over time.

This decoding step is crucial for identifying expletive subjects and clausal adjuncts, as well as for distinguishing accusative from dative objects.

In addition to the quantitative evaluation of the edge labeller delivered by the NEGRA/TIGER-derived test set, I also conducted a qualitative error analysis of the system's output. One finding from this work was that the edge labeller frequently fails to identify expletive subjects. For illustration, in the NEGRA/TIGER-derived SCF lexicon, `xa` is the most common SCF observed for the verb *geben* 'give' (in the sense of *es gibt* 'there is'); in contrast, in the output of the Berkeley Parser and edge labeller, `xa` is only the seventh most common SCF for *geben*, and the most common frame observed is `na` (transitive, with non-null subject). This reflects the problem that it is very difficult to distinguish expletives from ordinary neuter pronominal subjects, as there is usually little lexical evidence in the sentence to support one hypothesis over the other. Furthermore, the prior probability of an expletive subject is very low; in TIGER, expletives make up less than 1% of all pronouns. This shortcoming of the edge labeller is not repeated by the `mate-tools` parser; in the SCF lexicon from section 5.3, `xa` is the most common SCF tag for *geben* with 47.7% of instances, and `na` happens less frequently, with 20.5% of instances.

5.6.2 *Chronological evaluation*

Table 5.8 shows an evaluation of the SdeWaC-derived SCF lexicon over time using the various evaluations presented so far in this chapter; all these tests are conducted using the first 80M words of SdeWaC. The SiW2006 task from section 5.5.4 shows PairF and Rand_a evaluation values using both the Jensen-Shannon and skew divergence distance measures; the clusterings evaluated here are induced using the 'Random Mean' condition (averaging the results of ten trials). The SiWBrew2002 task from section 5.5.5 shows APP and MI values using the Jensen-Shannon distance measure; cluster initialisation is also random. The Duden task shows the F_1 value of the type-based intrinsic evaluation from section 5.4.

Syntactic analysis prior to 5.11.13 was delivered by the Stanford parser, and from that date on by the Berkeley parser. Because these two constituency parsers do not perform morphological analysis, the SCF lexica produced by them did not provide case information about PP complements. Intuitively, this information could help to distinguish, for example, locative from path PPs: compare accusative *in die Stadt* 'into the city' vs. dative *in der Stadt* 'inside the city'. The lack of this information might therefore reduce the performance on the automatic verb clustering evaluations. The final evaluation on 8.5.14 is conducted with the SCF tagger fed by the `mate-tools` dependency parser, where the subcategorisation acquisition system is no longer using the edge labeller.

5.7 CONCLUSIONS

This chapter has begun the empirical investigation of the argument structure of German verbs. In it, I have described the development of a SCF tagger and used it to derive a valency lexicon of German, recording the subcategorisation behaviour of verbs in a large corpus. I conducted an intrinsic evaluation with respectable results, and also an extrinsic evaluation with the task of automatic verb classification; I will cast the latter as a kind of standardised task and re-use it in subsequent chapters. We have seen empirical support for the hypothesis that verbs that exhibit a preference for entering into similar syntactic constructions may well belong to the same semantic verb class.

The extrinsic evaluation demonstrated that the SCF lexicon developed in this chapter is of better quality than the one developed by Schulte im Walde (2006), most likely due to its larger size. This volume of data leaves room to investigate whether verb descriptions can be made more detailed before data sparsity starts to become a serious problem; the following chapters will thus go in search of more fine-grained statistics on verb meaning.

The automatic verb classification experiments in this chapter have made use of several distance measures (Jensen-Shannon divergence and skew divergence), as well as several evaluation metrics for cluster purity (Rand_a, PairF, APP, and MI). We have seen that the performance of the clustering when using one distance measure highly correlates with the clustering performance using the other distance measures; likewise, I have shown that the cluster evaluation metrics are also highly correlated with each other. What this indicates is that the different measures are performing much the same job, and that there is really no need to use more than one distance measure, or more than one evaluation metric. Another observation made in the course of Experiment 1 was that hierarchical clustering with Ward's criterion performed better than k-means.

An early version of the SCF tagger and lexicon developed in this chapter, as well as Experiment 1 from the automatic verb classification task were published by Roberts, Egg and Kordoni (2014). The SdeWaC SCF lexicon was made available to the research community. To my knowledge, this is the first publicly available machine-readable subcategorisation lexicon of German. It has so far been used for a Bachelor's thesis (Zeller, 2018).

The Tgrep2 utility that I wrote for searching syntax trees was integrated into the Natural Language Toolkit (NLTK), a popular teaching and research library for NLP. It is available at `https://github.com/nltk/nltk/blob/develop/nltk/tgrep.py`.

As discussed in section 2.3.5, verb classes are commonly constructed on the basis of diathesis alternations of verbs. In this chapter, I also follow this approach, by representing verbs by their subcategorisation preferences. However, verb classes are also closely related to the notion of argument realisation, suggesting that verbs in similar classes will select similar arguments. As such, the endeavour of constructing verb classes shares much with the topic of frame semantics, and parallels can be drawn with the construction of other lexical semantic resources, such as FrameNet (Baker, Fillmore and Lowe, 1998) and PropBank (Palmer, Gildea and Kingsbury, 2005). The chapter after this one will delve into this topic in greater detail, where I investigate whether selectional preference information can also be used for automatic verb classification.

6

SELECTIONAL PREFERENCES

Chapter 5 explored automatic verb clustering based on verb subcategorisation preferences, replicating the method of Schulte im Walde (2006). A logical extension to this line of inquiry would be to further enrich the statistical description of verbs by adding information about what kinds of arguments tend to co-occur with verbs – the verbs' selectional preferences (cf. section 2.3.8).

Indeed, as I have outlined in section 4.3, there have been several studies that have investigated selectional preference modelling for this task, with varying results. Two previously published works, those of Schulte im Walde, and Sun and Korhonen, are especially relevant to this chapter. Recall that Schulte im Walde (2006) used a model of selectional preferences defined over top-level GermaNet synsets, and observed a small improvement in performance under certain conditions, but the improvements were not statistically significant, leading her to conclude that 'the 15 conceptual GermaNet top levels are not sufficient for all verbs' (p 189). This accords with the findings of Joanis (2002), who did a similar experiment for English, using a tree cut model on WordNet to model selectional preferences, and found that WordNet-derived selectional preference information hurt performance on a verb classification task. To my knowledge, no study has ever demonstrated that selectional preferences information is useful for automatic verb clustering in German.

In contrast, Sun and Korhonen (2009) found that selectional preference information did help substantially on automatic verb clustering in English, using a selectional preference model derived from corpus data with an unsupervised method.

In this chapter, I set out to explore in depth the question of whether selectional preference information can be helpful for automatic verb classification in German. In a single experiment, I shall compare a collection of methods for representing selectional preferences, including the methods of Schulte im Walde (section 6.2.4) and Sun and Korhonen (section 6.2.2). Several methods considered here have not previously been tested on automatic verb classification.

As the verb clustering experiments detailed in chapter 5 are an extrinsic evaluation of statistical descriptions of verbs, they can be seen as a kind of standardised test fixture. The two experiments in the last chapter delivered very similar results, and are run in much the same way; the most salient difference between them is that the gold standard dataset of the first experiment is larger than that of the second. Thus, the first experiment is used here with modification

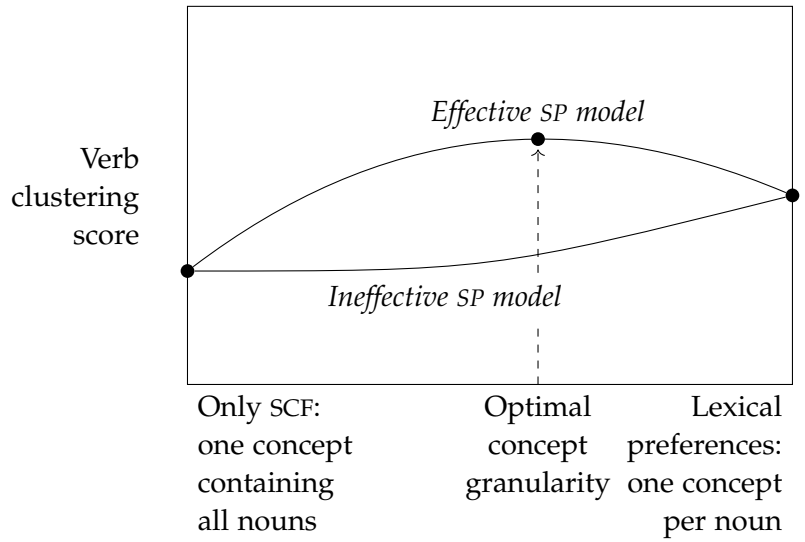


Figure 6.1: Granularity as a property of a selectional preference model.

to empirically compare different methods of selectional preference modelling. The approach I take is to 'parameterise' SCF tags with extra information which can describe the selectional preferences of a verb; this allows the joint distribution of SCF and selectional preferences to be modelled.

I propose to define a model of SP as some function that maps the lexical head of a verbal argument to some concept label. These concept labels are then appended to the SCF tag of a verb instance, always in a canonical order (subj, dobj, iobj, etc.). As an example, consider an instance of a verb with a transitive SCF (`na`), where the subject is mapped to the concept label `animate` and the object is mapped to `food`; the parameterised SCF tag would then be `na*subj-animate*dobj-food`.

The baseline SP model is a function that maps all nouns to the same concept label; this is isomorphic to having just SCF information. At the other end of the spectrum, each noun would be its own concept label; I term this a *lexical preference* model (section 6.2.1). Between these two extremes lie a continuum of other models, which map multiple nouns to the same concept label in some way. These models can be characterised both by their *granularity* – which can be defined to be, e. g., the mean number of nouns per concept label – as well as by the manner in which they group nouns together into particular concepts.

Figure 6.1 diagrams a principal hypothesis of this experiment: Some of these models of selectional preferences will be effective for verb clustering ('good' models), and some will be less effective ('bad' models). Intuitively, it seems as though the optimal SP model should fall somewhere in the middle of the spectrum, with concepts of moderate granularity that are put together in a way conducive to representing verbal argument preferences. The medium granularity of this hypothetical ideal model allows for generalisation to unseen data, mitigat-

ing against data sparsity, and also appeals to the naïve understanding of a noun class: that the verb *essen* 'eat' should tend to select a noun from the concept FOOD as its direct object.

Several assumptions underlie this experimental design:

1. That SP can be captured by attending solely to the lexical heads of the verb's arguments;
2. that nouns can be assigned to categories ('concepts') in a way that generalises over the selectional preferences of all kinds of verbs in the lexicon; and,
3. that these concepts will be equally useful for all kinds of grammatical relations (subject, object, prepositional object, etc.).

I shall use the results of the experiment to address the following questions:

1. What do good concept clusters look like? Are they similar to traditional semantic features such as +ANIMATE?
2. What is the right degree of granularity?

The next section introduces the experimental method employed in this chapter. In section 6.2, I describe the five models that are compared. Results of the experiment are presented in section 6.3 and are discussed in section 6.4. Finally, section 6.5 summarises the main conclusions of the chapter.

6.1 METHOD

In the following example:

(6.1) Wir benutzen Ihre Umfragedaten nicht für eigene Zwecke.
We use your survey data not for own purposes.
We will not use your survey responses for private purposes.

complements to the verb *benutzen* are the pronominal subject *wir*, the direct object *Umfragedatum* 'survey datum', and PP headed by *für* 'for' and with its argument *Zweck* 'purpose' in the accusative case. Recall that, under the baseline SCF-only model (cf. example 5.1), this combination of complements gave an SCF tag of `nap:für.Acc`.

For acquiring SP data, I only look at nouns, no pronouns. If an argument head to a verb is pronominal or not treated by a given SP model, it is not included in the SCF tag. For instance, the subject in example 6.1 is pronominal, and so is ignored for this verb instance.

Much prior work on selectional preferences has modelled granularity and concept space separately for each type of grammatical relation (Li and Abe, 1998; Clark and Weir, 2002); it is also common

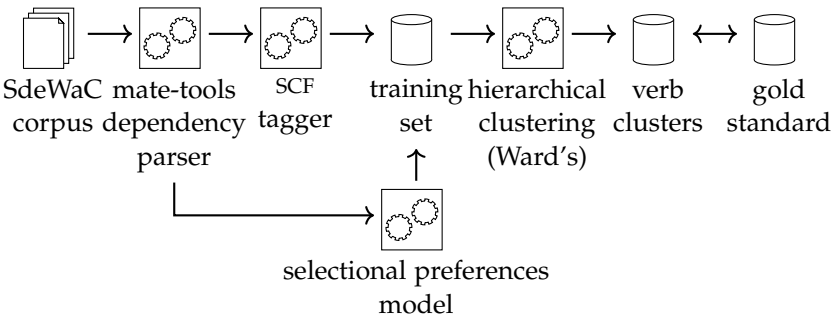


Figure 6.2: The experimental setup for comparing selectional preference models.

to model only particular grammatical relations or combinations of grammatical relations, as Schulte im Walde (2006) did (cf. section 4.3). In contrast, I use a simpler paradigm, where I treat every grammatical relation[1], and all grammatical relations are associated to the same concept space.

Under the simplest SP model, lexicalised preferences, discussed immediately below (section 6.2.1), each noun is mapped to itself. Under such a regime, the parameterised SCF tag would be

`nap:für.Acc*dobj-Umfragedatum*prep-Zweck`.[2]

With the categorised argument information added to the SCF tag, I now proceed using the exact same method described in section 5.5; the method is sketched in figure 6.2. For each verb instance in some collection, the *training set*, I compute the parameterised tag and count it. Each verb type is then represented by a vector over all these counts. For each verb, normalising its corresponding vector gives a discrete probability distribution over the possible lexical-syntactic frames that verb can occur in. I then use the Jensen-Shannon divergence to measure the dissimilarity between normalised vectors; as previously discussed in section 5.5.6, we can expect Jensen-Shannon to predict with high accuracy what the system's performance using the skew divergence would be, and so I do not employ the skew divergence here. The verbs are clustered using hierarchical clustering with Ward's criterion, and the automatically derived clustering is then compared to the gold standard, from which I can extract an evaluation measure.

A primary effect of adding the SP information in this way is to multiply the number of SCF frames, thus making the verb vectors longer. Since the same count information is spread over more dimensions, sparsity also increases.

1 As I will return to in section 6.3.2, I did actually test modelling using only certain grammatical relations or combinations of relations, but this did not improve results on this task.

2 *Zweck* is the 256th most common noun recorded in SdeWaC, so this would be captured by models with parameter $N \geq 256$. Actually, *Umfragedatum* not attested in SdeWaC, and so would not be captured in this example.

For this chapter, I use as a training set the first 3,000,000 instances of the 168 verbs contained in the verb clustering gold standard (cf. section 5.5.1). This requires about 270M words of parsed text, or about 25% of SdeWaC. There are 17,857 instances of each of these verbs, on average.

Again, as shown in section 5.5.6, it is sufficient to use only a single evaluation measure to judge the quality of a clustering. I use the PairF score, because the confusion matrix used to calculate it is useful for tests of statistical significance. As previously reported in section 5.5.4, the random baseline score (taken as the average score of 50 random partitions) is 2.08, and the optimal baseline is 95.81. I use only hierarchical clustering with Ward's criterion (Ward, Jr, 1963) and forgo k-means clustering, since Ward's performed better than k-means in the previous chapter. The hard clustering algorithm is used to partition the verbs in the test set into K groups, where K is again matched to the number of groups in the gold standard used for evaluation (43).

6.2 MODELS

6.2.1 *Lexical preferences*

This section introduces the *lexical preferences* (LP) model, the most fine-grained model of selectional preferences I shall admit here. Under LP, each noun represents its own concept; the concept mapping function is the identity transformation, mapping each noun to itself. The models described here are all controlled by one or more parameters. The LP model in this section includes only one parameter, N, which is the number of noun types included in the LP model, ordered from the most frequent noun type observed in SdeWaC to the least frequent. This follows the intuition that low-frequency nouns are unlikely to be useful for selectional preference modelling, as they increase data sparsity in a model and yet should seldom occur in test data.

The noun types in SdeWaC are counted and sorted by decreasing frequency. The top five noun types in the corpus are: *Jahr* 'year', *Mensch* 'person', *Zeit* 'time', *Kind* 'child', and *Land* 'country'. As mentioned, pronouns and noun types whose rank is greater than N are not added to SCF tags. Figure 6.3 shows how many of the training set's verb instances have one or more nouns appended to their SCF tags as a function of N. The apparent asymptote at 60% reflects the proportion of verbal argument instances in SdeWaC that are not nominal (i. e., pronouns or sentential complements). Hypothetically, the effect of increasing N should be to capture more information about verb-argument relations, albeit with diminishing marginal returns, indicated by the deceleration seen in figure 6.3 with high values of N. However, we should also expect higher N to increase model sparsity,

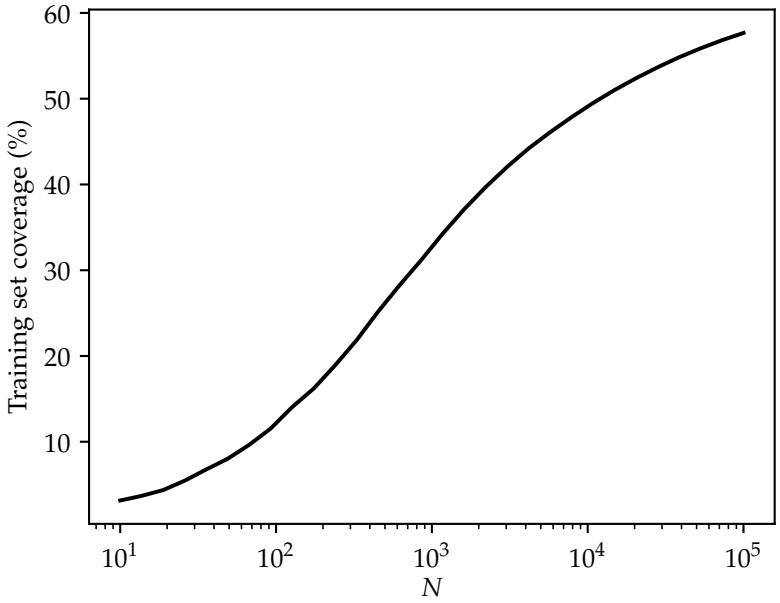


Figure 6.3: Fraction of verb instances in the training set parameterised by LP as a function of the number of nouns N included in the LP model.

which will frustrate automatic verb clustering. Therefore, I predict there should exist a 'sweet spot' for the value of N, neither too low nor too high. I explore the performance of this model using values of N taken from the set $\{5, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50000, 100000\}$.

6.2.2 *Sun and Korhonen*

Sun and Korhonen (2009) introduced a model of selectional preferences for automatic verb classification, which works by partitioning nouns into disjoint noun classes; the partition is induced from automatically parsed text in an unsupervised way. This method of acquiring and representing selectional preference concepts was also used by used by Shutova, Sun and Korhonen (2010) and Shutova, Teufel and Korhonen (2013) for metaphor identification.

Using the parsed SdeWaC, I record all nouns that are heads of arguments to some verb of interest, and compile a list of co-occurrence relations of the form $\langle \text{verb}, \text{grammatical-relation}, \text{noun} \rangle$. The grammatical relations considered are:

- subject;
- direct object (accusative); and
- indirect object (dative).

Count	Verb	Rel'n	Argument
69010	*spielen* 'play'	dobj	*Rolle* 'role'
52173	*sein* 'be'	subj	*Mensch* 'human'
30470	*sein*	subj	*Frage* 'question'
29768	*sein*	subj	*Kind* 'child'
29353	*werden* 'become'	subj	*Mensch*
28526	*haben* 'have'	subj	*Mensch*
27221	*sein*	subj	*Frau* 'woman'
25944	*haben*	dobj	*Recht* 'right'
23932	*sein*	subj	*Rede* 'speech'
23738	*werden*	subj	*Kind*
...			
20080	*geben* 'give'	dobj	*Möglichkeit* 'possibility'
16206	*stellen* 'put'	dobj	*Frage*
15862	*machen* 'make'	dobj	*Spaß* 'fun'
14405	*leisten* 'render'	dobj	*Beitrag* 'contribution'
13826	*stellen*	subj	*Frage*

Table 6.1: Most common verb-argument relations in SdeWaC.

I also experimented with recording two other kinds of grammatical relations:

- prepositional objects, where the grammatical relation records the preposition used. An example is $\langle$*geben*, prep-*in*, *Auftrag*$\rangle$, 'give' with PP using 'in' and prepositional argument head 'contract', meaning 'to commission sth.'; and
- adjectival modification of nouns (as in $\langle$*schwartz*, nmod, *Haar*$\rangle$, 'black' modifying 'hair');

however, these two relation types were not successful (cf. section 6.3.2).

I take as objects of interest the verb types listed in the SCF lexicon of section 5.3, retaining those noun-grammatical-relation-verb tuples that are seen 10 times or more in SdeWaC. This procedure gives observations on 60,870 noun types (representing 33,748,390 tokens); these noun types are associated with 11,426 verb-grammatical-relation types (6,705 verb types). An example is $\langle$*sprechen*, dobj, *Wort*$\rangle$, 'speak' with direct object 'word', which was observed 1,585 times. Table 6.1 shows the most common verb-argument combinations attested in SdeWaC. Much as we have already seen in table 5.2 (page 85), the table demonstrates that verbs such as *sein*, *werden* and *haben* are very common in the data, most likely because these can be auxiliaries. Note also that

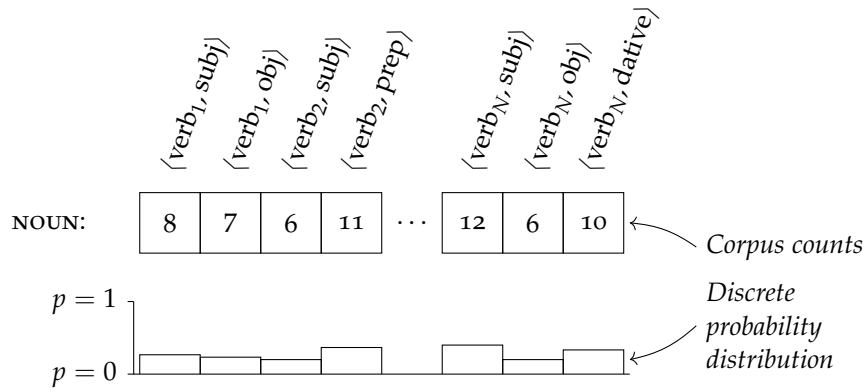


Figure 6.4: Nouns as vectors of discrete probabilities over possible verb-grammatical-relation combinations.

the subject relation is very common, as I have previously noted in section 5.6.1.

I now proceed using the same method as the one I use for clustering verbs in chapter 5 and in this chapter. It is diagrammed in figure 6.4; please note the similarity to figure 5.3 (page 91). Each noun is represented by a 11426-dimensional vector recording the noun's co-occurrence counts with the various verbs and grammatical relations. These vectors are then normalised, giving for each noun a discrete probability distribution over the set of verb-grammatical-relations. These vectors are then automatically clustered into M disjoint groups using hierarchical clustering with Ward's criterion, with the Jensen-Shannon divergence (equation (5.2)) used as the distance measure.[3]

The concepts that result from this process do not have names; examples from the clustering can be seen in table 6.4. Noun arguments to a verb are mapped to the concept that contains that noun; for example, all nouns in the first cluster will be labelled as belonging to `concept1`.

The parameter M is a convenient way to control the granularity of the SUN model. Exactly like with the LP model, I also use the parameter N to indicate how many of the most common noun types in SdeWaC are included in the model. I search the parameter space $N \in \{300, 500, 1000, 5000, 10000\}$ and $\frac{N}{M} \in \{5, 10, 15, 20, 30, 50\}$.

3 Note that Sun and Korhonen (2009) partition their nouns using spectral clustering with the MNCut algorithm (Meilă and Shi, 2001); in my experiments, spectral clustering was almost identical in performance to hierarchical clustering, but computationally more expensive, which is why I have chosen to use hierarchical clustering here. I do make use of MNCut for the word space model described in the next section (section 6.2.3).

6.2.3 *Word space model*

This section introduces a model of selectional preferences that also works like the SUN model by creating a partition over nouns. In this case, I induce a word space model (cf. section 3.2) for German nouns, without the use of syntactic analysis, and then use the WSM's pairwise similarity scores to cluster the nouns into disjoint sets.

Word co-occurrence counts are computed on the SdeWaC corpus, with features being the 50,000 most common words in SdeWaC, setting aside the first 50 most common words (which are assumed to be stop words); that is, the words with rank 51–50,050 inclusive.

Sentences are taken as word co-occurrence contexts, the corpus is lemmatised, and punctuation is removed. Co-occurrence counts are weighted using the t-test scheme: For a given word w_i and feature c_j,

$$\text{ttest}(w_i, c_j) = \frac{p(w_i, c_j) - p(w_i)p(c_j)}{\sqrt{p(w_i)p(c_j)}} \tag{6.1}$$

After this, I apply a technique called *context selection* (Polajnar and Clark, 2014), a kind of regularisation or smoothing. Each word vector is made sparser by setting all dimensions to zero, except for the C dimensions with the largest absolute values in the vector. The value of C was tuned by optimising the correlation between the WSM model's semantic similarity scores and human manual judgements. The data set used for this was collected by Gurevych and Niederlich (2005), who solicited human similarity judgements for 65 word pairs.

Maximum performance is obtained with $C = 380$, giving Spearman $\rho = 0.813$ and Pearson $r = 0.707$ correlations against the mean human-assigned similarity score. For comparison, on this task, the human inter-annotator agreement was measured as $r = 0.810$.

Using the cosine similarity measure, the WSM estimates similarities between all nouns in SdeWaC. After this, the top N most common nouns can be selected, and the similarity scores arranged into a symmetrical similarity matrix of shape $N \times N$. This similarity matrix is then used to partition the N nouns into M disjoint sets; for this I use spectral clustering with the MNCut algorithm (Meilă and Shi, 2001). The resulting clustering or noun partition is used the same way as the SUN model. I search the same parameter space as for the SUN model.

6.2.4 *GermaNet*

In this section, I introduce a generalisation of the model used by Schulte im Walde (2006). She used the 15 top-level GermaNet syn-

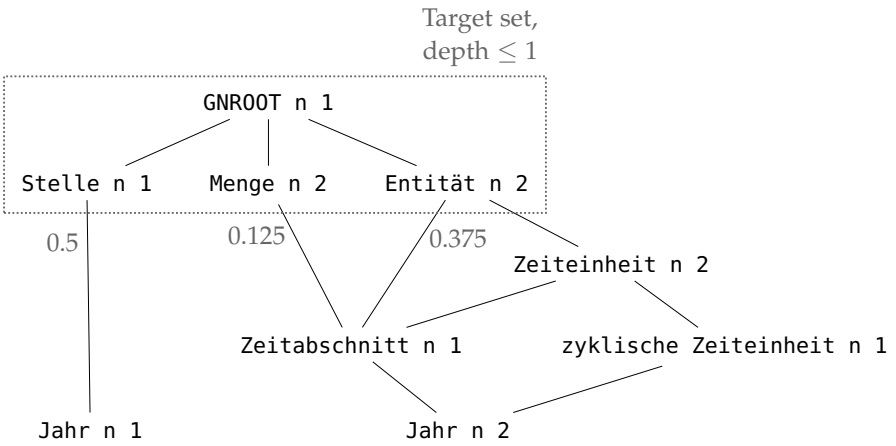


Figure 6.5: The target sets of nouns in the GermaNet SP model.

sets (section 3.3) as a concept inventory[4]; I will refer to this collection of synsets as the *target set*. I generalise the model by defining the target set to be the set of synsets in the GermaNet noun hierarchy that are at a particular depth d or less:

$$\{s \mid \text{depth}(s) \leq d\}$$

where the depth of a given synset s is defined to be the smallest possible number of hyponymy links connecting s to the root node `GNROOT` of the hierarchy.

There are 6 synsets in the target set with $d = 1$ (`GNROOT`, and its five daughter synsets `Entität n 2` 'entity', `Menge n 2` 'quantity', `Spezifikum n 1` 'attribute', `Stelle n 1` 'location', and `Zustand n 1` 'condition'). There are 17,125 synsets with $d = 6$; the maximum depth of the GermaNet hierarchy is 21.

A given noun argument to a verb is assigned to one or more representative synsets in the target set in the following way. If a noun is already in the target set, it is assigned its own synset as its concept label. If it is not already in the target set, its concept label is taken to be its lowest hypernym that is in the target set. Polysemous nouns vote for which element of the target set represents them, with one vote cast by each path from a synset to a member of the target set. These votes are then normalised, giving a set of weights over one or more members of the target set.

Figure 6.5 shows the concept labelling from the GermaNet model with parameter $d = 1$ for the noun *Jahr* 'year', showing how the noun, taking into account its various senses, is mapped to a fuzzy set of the top-level synsets. The displayed hypernym hierarchy has been simplified, and is actually deeper than pictured here.

4 As GermaNet was not complete at the time, she reports that some of these 15 top-level synsets were 'manually added' (p 168); it is not clear which synsets were added and how.

Note that the fact that a given noun can map to more than one concept label (synset in the target set) means that this model of selectional preferences uses a soft clustering (cf. section 3.6), where the notion of a noun partition is relaxed, and nouns can be said to belong to varying degrees to multiple concepts. As we shall see, soft clustering models produce much longer verb vectors.

For the GermaNet model, I search the parameter space of $d \in \{1, \ldots, 8\}$. There is no equivalent to N in this model. Rather, all nouns found in GermaNet are always associated with some concept label(s); only the number of possible concept labels changes with varying values of d.

6.2.5 *Latent Dirichlet allocation*

Latent Dirichlet allocation (LDA: Blei, Ng and Jordan, 2003) is a generative model used to discover structure inherent in some unlabelled data. In NLP, LDA is often used to build *topic models*, the underlying assumption being that a given document will be written about a small number of possible topics, and that each topic is associated with a certain vocabulary. This intuition is formalised in a probabilistic graphical model. Given a parameter K for the desired number of topics, LDA imputes a series of latent variables (e. g., the word probabilities for a given topic z); the model then ingests raw text and uses statistical inference to estimate these probability distributions.

This section introduces the 'LDA' selectional preferences model of Ó Séaghdha (2010), who extends the topic model approach to model the relationship between predicates and their arguments (specifically the verb-object relation), as mediated by latent variables representing different kinds (classes) of argument. I have used only his simplest model, and extended it to accommodate more than a single grammatical relation at a time.

The generative story behind this model is diagrammed in figure 6.6 in plate notation. For a given verb v, and for a given grammatical relation r that the verb has in its set G, I sample a noun class z from a multinomial distribution $\Phi_{v,r}$ with a Dirichlet prior parameterised by α; there are K total topics or 'noun classes' that I will take to be the SP concept inventory. I then sample a noun n from a multinomial distribution Θ_z with a Dirichlet prior parameterised by β. I repeat these steps until S words have been generated for this $\langle \text{verb}, \text{grammatical relation} \rangle$ pair. Again, the bag of nouns W so generated are the only observable variables of the whole process; the other nodes are latent variables. Like Ó Séaghdha, I use an asymmetric Dirichlet prior for $\Phi_{v,r}$ (i. e., α can differ for each noun class), and a symmetric prior for Θ_z (β is the same for each Θ_z).

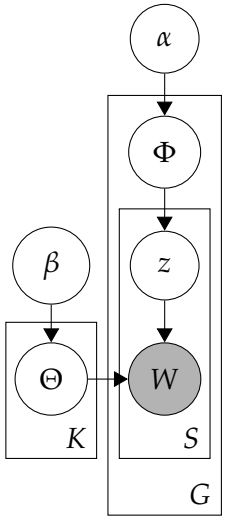


Figure 6.6: Plate notation of the Latent Dirichlet allocation SP model.

I construct the model using MALLET (McCallum, 2002)[5], training the model with the same co-occurrence statistics as were used for building the SUN model. The model is trained for 1,000 iterations using the software's default parameters, and the hyperparameters α and β are re-estimated every 10 iterations. To model concept granularity, I construct models with $K = 50$ or $K = 100$ topics. As before, I use the parameter N, searching the space $N \in \{500, 1000, 5000, 10000, 50000, 100000\}$.

Using this recipe, LDA creates a soft clustering of nouns, like I have in the GermaNet SP model. That is, a given noun will belong to one or more concepts (noun classes), with varying degrees of strength. Hypothetically, this seems like an attractive paradigm, since the ability of a noun to belong to multiple concepts could be a good way to model the polysemy of nouns. For completeness, I also test a hard clustering version of the LDA model. Here, each noun n is mapped to the label of the noun class z it belongs most strongly to (the cluster with the highest weight); I query the model to give the most likely class label

$$\arg\max_{z} P(z|n)$$

6.3 RESULTS

6.3.1 *Major effects*

Table 6.2 summarises the performance of the various SP models examined in this chapter, showing the best parameter settings found for each model in ranked order. Also displayed are the number of SCF types produced by the model, equivalent to the length of the verb vec-

5 http://mallet.cs.umass.edu/

SP model	Parameters	Granularity	PairF	Number of SCF types
SUN	10,000 nouns	1,000 noun classes	39.76	248665
LDA (hard)	10,000 nouns	50 topics	39.10	78409
LP	5,000 nouns		38.02	388691
WSM	10,000 nouns	500 noun classes	36.95	149797
LDA (soft)	10,000 nouns	50 topics	35.91	1524338
GermaNet (soft)	depth = 5	8,196 synsets	34.41	851265
GermaNet (hard)	depth = 8	10,900 synsets	34.04	428043
Baseline			33.47	673

Table 6.2: Summary of evaluation results for selectional preference experiments.

tors; as explained in section 6.1, the baseline method's 673 different SCF types are multiplied out as a function of the number of concepts in a particular SP model, increasing data sparsity.

Taking the confusion matrices underlying the F_1 scores attained by two given models, it is possible to calculate the statistical significance of a difference in scores using McNemar's test. I find that all models except GermaNet-hard perform better than the baseline at at least the $p < 0.01$ level; the LDA-hard model is better than the GermaNet, LDA-soft, WSM and LP models at at least the $p < 0.05$ level; and the SUN model is better than all models except LDA-hard at at least the $p < 0.05$ level. All other differences in score are not statistically significant.

6.3.2 *Minor effects*

A series of somewhat ad-hoc and non-exhaustive searches were performed on the parameter spaces of the various SP models. By locating paired data points within this set of results, it is possible to experimentally measure the effect of several parameters of the procedure on the clustering performance.

As stated in section 6.1, I model selectional preferences on all grammatical relation types (subject `subj`, accusative object `dobj`, dative object `iobj`, prepositional object `pobj`). While conducting the experiment, I measured the effect of only modelling certain relations or combinations of relations; specifically, I tried `all`; `subj`; `obj`; `subj+obj`; `iobj`; `pobj`; `obj+iobj`; and `obj+iobj+pobj`. In theory, certain grammatical relations might be more amenable to modelling; however, in practice, filtering out grammatical relation types reduced the *coverage* of the training set (the fraction of the training set that was parameterised in some way for SP information) and increased data sparsity. My best results were obtained using all grammatical relations. This finding stands in contrast to the conclusions of Sun and Korhonen (2009), whose SP model performed best using only the subject relation, as

well as Schulte im Walde (2006), who obtained her best results using only specific combinations of SCF-grammatical-relation slots.

I also tested whether it was better to include PP case information in SCF tags, as has been my standard procedure since section 5.1, or to strip out this information. Theoretically, leaving out PP case information from the SCF tags would result in fewer SCF types and hence shorter verb vectors, and this would tend to reduce data sparsity. However, the experimental results showed a significant reduction in performance without PP case information across all models. The baseline model also scores 4% lower in this condition (PairF = 29.70).

As mentioned in section 6.2.2, I tested whether it was better to include noun-adjective relations in the grammatical-relation-based noun clustering methods (SUN and LDA), finding that including these relations produced worse performance.

I also experimented with allowing pronouns to be used as arguments to verb instances. I tried including pronouns under two conditions: including only first and second person pronouns; and including all pronouns. All models performed significantly better when pronouns were not included as arguments to be categorised.

Additionally, the following parameters were explored for the SUN model:

- Whether to include prepositional relations: if enabled, the grammatical relation data are also collected on nouns which are prepositional objects of verbs; the resulting relation tuple would look like $\langle \text{verb}, \text{prep-}um, \text{noun} \rangle$. The best results were obtained by not including prepositional concomitants.

- Which distance measure to use for clustering the nouns: I experimented with Jensen-Shannon divergence, skew divergence, and cosine distance. Jensen-Shannon performed significantly better than the other two measures.

- Whether to re-weight the verb-relation-noun counts before clustering: I tried re-weighting the corpus counts using pointwise mutual information and t-test schemes (equation (6.1)). Despite some promising initial results from the t-test, I ultimately chose not to perform re-weighting.

- Whether to include auxiliary and modal verbs in the grammatical relation data: This results in additional relation tuples in the data, with verbs like *werden* 'become' and *haben* 'have'. Including these verbs decreased performance.

A final analysis was performed by regressing the N parameter and the various combinations of grammatical relations, which determine the training set coverage, or number of verb instances parameterised for selectional preferences, against the verb clustering performance.

For all SP models except GermaNet, I found a positive correlation between the number of verb instances parameterised and the PairF score on the verb clustering task; this correlation was persistent and showed up regardless of other parameters and their settings (e. g., granularity, etc.). That is, the more training set data is captured by the SP model, the better the performance attainable on the verb clustering task; this effect appears for all models but GermaNet and is independent of the model's internal constitution. This observation supports the conclusion that all these models are effective for capturing SP information.

6.4 DISCUSSION

6.4.1 *Comparing models*

The two GermaNet models are the least successful of the group, performing only slightly better than the baseline. The soft clustering model comes off better than the hard clustering version, but the difference is not very large. Both models perform best with intermediate values of d. This accords with the report from Schulte im Walde (2006) that top-level GermaNet classes are not adequate for modelling selectional preference concepts.

The hard clustering LDA model performs much better than the soft clustering, attaining the second highest score in this evaluation. The soft LDA model does not seem to capture polysemy as might have been hoped; it also produces verb vectors more than an order of magnitude longer than the hard clustering version, which presumably aggravates problems with data sparsity. Both versions of LDA perform better with 50 noun classes than with 100.

Of the topics constructed by LDA, some can be assigned a fairly cohesive label; for example, body parts, people, quantities, emotions, places, buildings, tools, etc. Other topics seem less narrowly focused. In particular, it appears that high frequency words are often generated with high probability by several topics, although these topics do not seem to be capturing different word senses. The $K = 100$ models demonstrate this tendency to a greater degree. An example is *Zeit* 'time', which strongly belongs to three topics in the $K = 50$ models. Only the highest ranked of these is actually a collection of time expressions, and the other two are topics with many semantically unrelated words. In the $K = 100$ models, *Zeit* belongs to six topics; again, only the topic with the highest value of α contains other time expressions. In the $K = 50$ models, there are 11 topics to which I am unable to assign a clear label; in the $K = 100$ models, this number is 38. Thus, from this manual examination of the produced LDA topics, it seems that higher values of K do not produce more semantically specific noun groupings.

α	Members
0.404	*Mann* 'man', *Frau* 'woman', *Gott* 'God', *Herr* 'Mr.', *Vater* 'father', *Leute* 'people', *Mutter* 'mother', *Freund* 'friend', *Jesus* 'Jesus', *Eltern* 'parents', *Junge* 'boy', *König* 'king', *Sohn* 'son', *Mädchen* 'girl', *Bruder* 'brother', *Tochter* 'daughter', *Freundin* 'girlfriend', *Schwester* 'sister', *Spieler* 'player'
0.250	*Mensch* 'human', *Kind* 'child', *Frau* 'woman', *Mann* 'man', *Tier* 'animal', *Leute* 'people', *Person* 'person', *Familie* 'family', *Million* 'million', *Hund* 'dog', *Patient* 'patient', *Prozent* 'percent', *Soldat* 'soldier', *Mädchen* 'girl', *Jude* 'Jew', *Sohn* 'son', *Junge* 'boy', *Gruppe* 'group', *Volk* 'people'
0.212	*Hand* 'hand', *Kopf* 'head', *Herz* 'heart', *Körper* 'body', *Fuß* 'foot', *Haut* 'skin', *Baum* 'tree', *Haar* 'hair', *Boden* 'ground', *Arm* 'arm', *Bein* 'leg', *Erde* 'earth', *Zahn* 'tooth', *Finger* 'finger', *Loch* 'hole', *Gesicht* 'face', *Blatt* 'leaf', *Auge* 'eye', *Pflanze* 'plant'
0.202	*Welt* 'world', *Mensch* 'human', *Leben* 'life', *Land* 'country', *Gesellschaft* 'society', *Staat* 'state', *Kultur* 'culture', *Stadt* 'city', *System* 'system', *Art* 'kind', *Markt* 'market', *Kirche* 'church', *Gruppe* 'group', *Geist* 'spirit', *Natur* 'nature', *Wirtschaft* 'economy', *Familie* 'family', *Körper* 'body', *Europa* 'Europe'
0.159	*Regierung* 'government', *Staat* 'state', *Land* 'country', *Deutschland* 'Germany', *Unternehmen* 'company', *USA* 'United States', *Bundesregierung* 'Federal government', *Partei* 'Political party', *Stadt* 'city', *SPD* 'SPD', *Union* 'union', *Kirche* 'church', *Israel* 'Israel', *PDS* 'PDS', *EU* 'EU', *CDU* 'CDU', *Firma* 'company', *Gewerkschaft* 'labor union', *Politik* 'politics'

Table 6.3: Highest-weighted topics in the LDA SP model with $K = 50$.

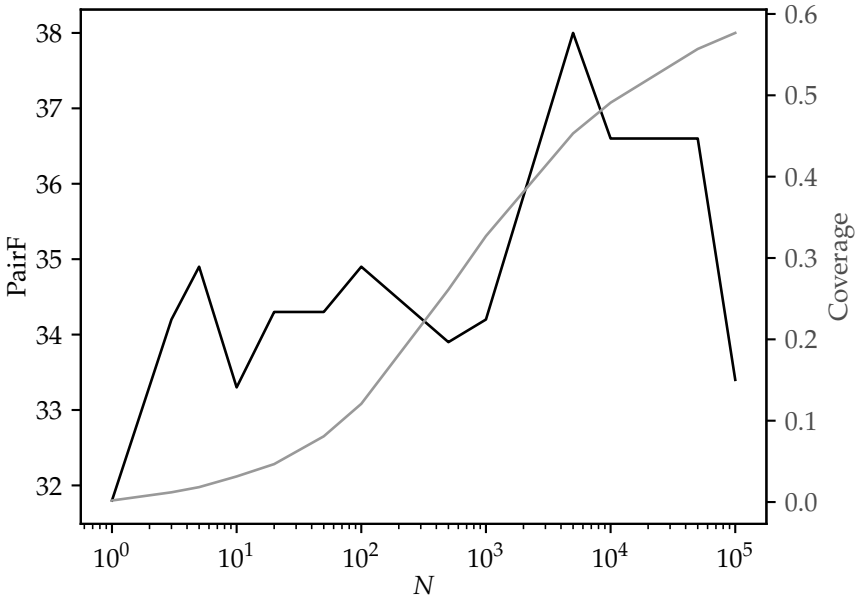


Figure 6.7: Verb clustering performance (black) and training set coverage (grey) of the LP model as a function of the number of nouns N included in the model.

For illustration, table 6.3 shows the highest-weighted noun classes from the $K = 50$ LDA model. The two top clusters are both focused on kinds of people, and exhibit considerable overlap. It is difficult to guess what distinction the model is drawing between these two groups; perhaps the first grouping is more agentive and the second grouping less so. One wonders whether the large intersection between these two groups may not be counterproductive for modelling SP. The third grouping is largely about body parts, but also appears to include kinds of plants. The fourth grouping contains social constructs and spiritual drives, but also generates the words 'human' and 'family', in common with the second group. The fifth grouping contains a number of collective nouns referring to kinds of organisations, likely a useful abstraction.

The LP model is very effective, which is perhaps surprising, given its extreme simplicity. As hypothesised, larger values of N eventually lead to sparsely problems, as can be seen in figure 6.7. We can see that optimal performance on this training set is obtained with values of N between 1,000 and 50,000.

The best performing model in this evaluation is the SUN model; it attains its best performance with parameters $N = 10000$ (the maximum value I tested), and $M = 1000$. This represents relatively fine grained clusters, with average concept size $\frac{N}{M} = 10$. This accords with the findings of Sun and Korhonen (2009), who report that their best results were obtained with 10–16 nouns per cluster on average.

LKW 'truck', *Lkw* 'truck', *Lastwagen* 'truck', *Castor* 'container for highly radioactive material', *Laster* 'truck', *Krankenwagen* 'ambulance', *Transporter* 'van', *Traktor* 'tractor'
Hand 'hand', *Kopf* 'head', *Fuß* 'foot', *Haar* 'hair', *Bein* 'leg', *Arm* 'arm', *Zahn* 'tooth', *Fell* 'fur'
Leiche 'corpse', *Leichnam* 'body', *Schädel* 'skull', *Skelett* 'skeleton', *Wrack* 'wreck', *Mumie* 'mummy', *Trümmer* 'debris'
Sauna 'sauna', *Badezimmer* 'bathroom', *Schwimmbad* 'swimming pool', *Nachbildung* 'replica', *Kamin* 'fireplace', *Aufenthaltsraum* 'common room', *Mensa* 'cafeteria'
Rechnung 'bill', *Kopftuch* 'headscarf', *Uniform* 'uniform', *Anzug* 'suit', *Helm* 'helmet', *Gewand* 'garment', *Handschuh* 'glove', *Mitverantwortung* 'joint responsibility', *Bart* 'beard', *Rüstung* 'armour', *Mitschuld* 'complicity', *Socke* 'sock', *Jeans* 'jeans', *Sonnenbrille* 'sunglasses', *Aufschrift* 'inscription', *Pullover* 'sweater', *Weste* 'vest', *Handschellen* 'handcuffs', *Hörner* 'horns', *Kennzeichen* 'marking', *Tracht* 'traditional costume', *Korsett* 'corset', *Schuhwerk* 'footwear', *Kopfbedeckung* 'headgear', *Pelz* 'fur', *Maulkorb* 'muzzle'
Missionar 'missionary', *Weihnachtsmann* 'Santa Claus', *Selbstmordattentäter* 'suicide bomber', *Bote* 'messenger', *Nikolaus* 'Nicholas', *Killer* 'killer', *Bomber* 'bomber', *Osterhase* 'Easter bunny'

Table 6.4: Example noun clusters in the SUN SP model.

Some of the concept clusters derived by the SUN model are shown in table 6.4. These examples make obvious that concept clusters are often organised as collections of synonyms or co-hyponyms, and often include alternate spellings, as can be seen in the 'truck' grouping. The grouping with articles of clothing also includes the nouns 'bill', 'joint responsibility', and 'inscription', all of which can be 'worn' or 'borne' in German (*tragen*). Other groups are more thematically related, as with the cluster containing 'corpse' and 'body'. All month names are grouped together in one 12-word cluster. Some clusters seem to represent combinations of concepts; for example, the model groups together sports, musical instruments, and dramatic roles, these all being things that can be played (*spielen*). Many clusters are devoted to groups of names or proper nouns, often of a very specific nature. Examples are: professional roles, such as titles of government ministers; names of diseases and medications; geographical locations, such as the names of Eastern European countries; and many collections of proper names, such as a cluster of author names, Biblical names, philosophers, NGOs, foreign currencies, German male first names, newspapers, or television channels. The last grouping in table 6.4 combines two of these specific clusters, lumping Santa Claus and the Easter Bunny in with killers and suicide bombers.

6.4.2 *Noun classes as concepts*

The WSM model does not perform nearly as well as the SUN model. Since both the WSM model and the SUN model employ the same approach to representing selectional preference concepts, namely a clustering of nouns, this affords an opportunity to qualitatively compare an effective clustering against a less effective clustering. Note that the WSM model's partition of nouns is based on paradigmatic information, meaning that a noun is clustered based on which sentence contexts it appears in. By contrast, the SUN model operates on syntagmatic information, whereby a noun is clustered based on which grammatical contexts it appears in. Therefore it is not very surprising to observe that the WSM clusters are almost all organised thematically, and the synonym/co-hyponym structure characteristic of the SUN classes is missing. A representative example is {*Pferd* 'horse', *Reiter* 'rider', *Stall* 'stall', *Sattel* 'saddle', *Stute* 'mare'}. Here we can see that WSM has mixed together several thematic roles in a single grouping: While a rider is probably an Agent, the horse should be a Theme or Patient, and the stall is a Location. This conflation of roles into one concept likely makes the WSM model less effective for modelling selectional preferences.

Another more quantitative difference between the WSM and SUN models can be seen through an examination of the distribution of cluster sizes in both models. Figure 6.8 displays histograms of the

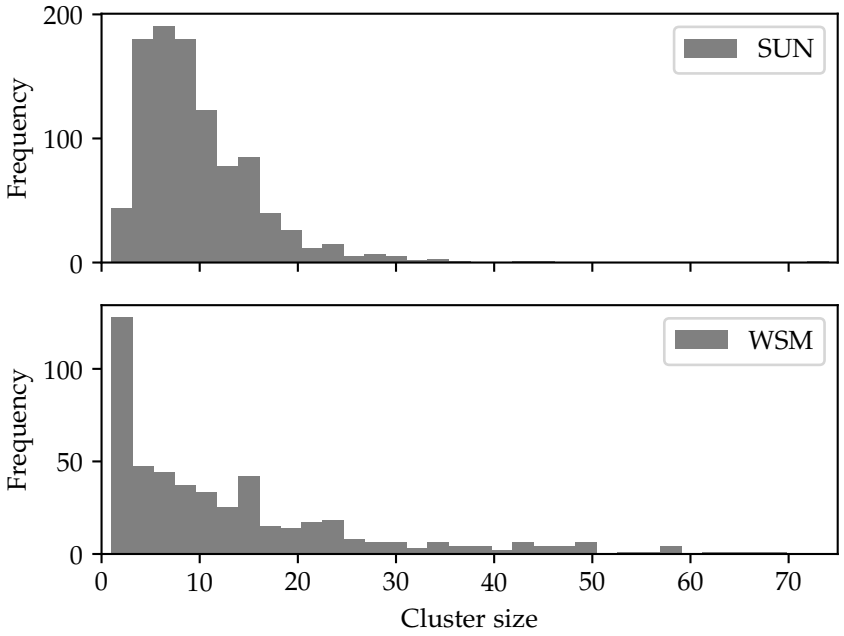


Figure 6.8: Histograms of noun cluster sizes for the SUN and WSM models.

noun cluster sizes for both models. Note that both models here include the same number of nouns ($N = 10000$); even taking into account the larger average cluster size of the WSM model (with $M = 500$) compared to the SUN model (with $M = 1000$), the differences in cluster sizes between the two models are still readily apparent.

The WSM model has many small noun clusters and a long tail of very large noun clusters (cluster size variance 2,800). The model constructs 56 singleton clusters. Its largest cluster contains 1,076 high frequency nouns that are semantically unrelated ('day', 'question', 'case', 'part', 'reason', 'kind', 'week', 'person', 'month', 'group', 'interest', etc.). These nouns are likely classified together because they are semantically neutral, in the sense that they appear in many contexts; this one cluster by itself covers 13.67% of all noun tokens in SdeWaC.

By contrast, the SUN model has much more tightly grouped cluster sizes (cluster size variance 37). The smallest classes constructed by SUN contain two nouns each; there are 12 such noun pair classes. The largest class in the model contains 73 nouns, including 'gas', 'taboo', 'pioneer', 'mustard', 'spy', 'mafia', and 'skinhead'. While these do appear to be unrelated words, they only account for 0.1% of the tokens in SdeWaC. The next two largest classes in SUN are semantically coherent; both with 40 nouns, they comprise politicians' surnames and male first names. The most common noun class in SUN contains 28 nouns and represents 3.6% of the noun tokens in SdeWaC; this group-

ing also looks semantically acceptable, containing 'human', 'child', 'woman', 'man', 'people', 'Mr.', 'mother', and 'father'.

Considering the conflation of thematic roles in the WSM classes, as well as the distribution of cluster sizes such that many nouns end up in a few very large clusters that do not select for specific contexts, it is reasonable to wonder why the WSM model is effective in this evaluation at all. It appears that, even though they are organised thematically, many clusters in the WSM model collect together related terms and thus probably do represent useful abstractions. Examples are clusters containing body parts, country names (there are separate classes for European, African, Asian, etc. countries), disease names, human names, articles of clothing, and the cluster {'fruit', 'apple', 'banana', 'pear', 'strawberry'}.

6.4.3 *Effect of training set size*

The comparative success of the lexical preference model raises the question of whether this model performs particularly well on this task because there is not enough training data for the other models to achieve better scores.

The accuracy of the subcategorisation preference vectors estimated by any given SP model is determined by the size of the training set, and I assume that this accuracy will be reflected in the PairF statistic when the model is evaluated. The size of any given model (i. e., the verb vector length) is a function of that model's parameters (N, K, d, etc.). The size of the training set and the size of the model together influence the degree of data sparsity in the model, which I have also previously mentioned as a potential issue that could negatively impact the PairF.

To get a closer picture of the effects of data sparsity on model performance, I conducted a second experiment. Here I took the four best performing models from the first experiment and tested them, varying the training set size as an independent variable from 10,000 verb instances all the way up to the full size of SdeWaC (11 million verb instances).

The results of this inquiry are shown in figure 6.9. We observe that sparsity is a problem for all models when the training set contains fewer than 3×10^5 verb instances; in this domain, the baseline method performs best. Above this threshold, the training set size is large enough to create models that work better than the baseline. The LP model's evaluation results are consistently good, and neatly follow the results of the SUN model. The LDA-hard model has only 50 topics, and seems to do better than the other larger models with fewer data, though it appears to become less competitive as the training set size increases above three million verb instances. At the maximum training set size, the best performing models are LP, SUN, and WSM. The

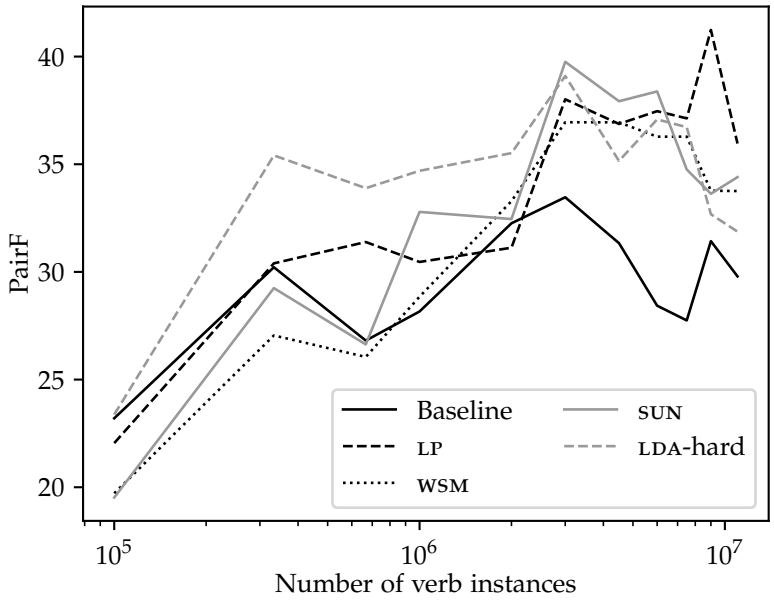


Figure 6.9: Verb clustering performance of SP models as a function of number of verb instances.

figure also illustrates a property of the evaluation performed here: The plotted curves frequently look jagged or wiggly. This reflects the instability of the hierarchical clustering algorithm, which is both deterministic and greedy; early clustering mistakes can compound over time, so that two similar inputs to the algorithm can produce two quite different output clusterings.

6.5 CONCLUSIONS

This chapter has delved deeper into empirically modelling the argument structure of German verbs by presenting an experiment that directly compared five different models of selectional preferences. We have seen that selectional preference information can be helpful for verb classification; all of the models tested here perform better than the baseline method. This finding is in agreement with Sun and Korhonen (2009), and stands in contrast to the results of Schulte im Walde (2006) and Joanis (2002). I believe this study to be the first empirical result demonstrating that SP information can improve performance on automatic verb classification in German.

Having now analysed the results, it is time to revisit the questions posed in chapter's introduction. So: what do good concept clusters look like? And: what is the right degree of cluster granularity?

The results suggest that noun classes function best when they are relatively small and semantically cohesive. Intuitively, groups of syn-

onyms and co-hyponyms should be helpful for modelling verbal selectional preferences; this represents a clear case where a SP model of intermediate granularity can make effective use of generalisations that are not possible in the LP model. Other groupings that are useful for modelling selectional preferences include classifications of proper names, for collecting together entities of various types: human, corporate, geographical, or kinds of medications. Furthermore, strong models such as SUN do seem to organise concepts in a way that is consistent with respect to theoretical semantic features such as animacy, concreteness, abstractness, and so on.

This study has also demonstrated that smaller noun clusters seem to perform better than large clusters. As we have seen in section 6.4.3, there is a trade-off to be made between the size of the training set and the granularity of the clusters that is mediated by data sparsity.

We have also seen that the LDA-hard and SUN models often work better than LP, suggesting that the optimal level of granularity is greater than one noun per concept. However, on a big picture level, the LP model is very simple to implement, and is only slightly worse than SUN and LDA, which are both considerably more complicated.

Finally, the comparison between the WSM and SUN models suggests that syntagmatic knowledge is more useful for modelling selectional preferences for automatic verb classification than paradigmatic knowledge.

The major results of the experiments performed in this chapter were published by Roberts and Egg (2014); this article offered the first empirical comparison of SP models for automatic verb classification. The computer code that I wrote to use the GermaNet lexicon has been made freely available under the name `pygermanet`[6].

This chapter set out to test whether selectional preferences can be a useful source of information for verb classification. I have demonstrated that the combination of syntactic frame information and lexical information about arguments, where applicable, outperforms clustering based on syntactic frame information alone. I submit that the corpus counts of co-occurrences between verbs and SCFs is a way of directly modelling the argument structure of verbs. By parameterising the SCFs for selectional preference information, the argument structure can be captured in more detail.

Having thus considered argument structure, I will move on to to the other facet of verbal lexical semantics, namely aspectual structure, the category that concerns the temporal structure of a verbal event. While argument structure and aspectual structure may be assumed to overlap, theory predicts that they should be fundamentally different categories. As we shall seen in the next chapter, aspectual structure is to a large degree not immediately ascertainable from the raw textual

6 `https://github.com/wroberts/pygermanet`

input, unlike argument structure. Thus, progress in this direction will necessitate the creation of a manually annotated resource.

Part III

ASPECTUAL STRUCTURE

7

ASPECTUALLY ANNOTATED VERBS

This chapter switches tack and embarks in a direction that I have so far not explored in this thesis, aiming to model the aspectual structure of German verbs. Section 2.4 has already introduced and motivated the category of lexical aspect on a theoretical level; and section 4.4 has surveyed prior work on computational lexical aspect. Here, I get my hands dirty with an empirical exploration of the aspectual features and classes of the verbs in the SdeWaC corpus.

I begin in section 7.1 by offering some notes on aspectual vagueness and ambiguity, a problem identified in semantic theory (cf. section 2.4.8), but one that is not always treated in applied methods.

A second complication is that aspectual structure cannot be automatically extracted from a syntactic analysis of a sentence in the way that argument structure can be. Siegel and McKeown, working in English, have attempted to circumvent this problem by working with linguistic indicators, syntactic structures and modifiers that correlate to some degree with lexical aspect. In section 7.2, I attempt this as well, translating the linguistic indicators of Siegel and McKeown into German, and evaluate. Because the results of this are not very promising, I produced a corpus of German verb instances manually annotated for aspectual features; this work is presented in section 7.3.

Section 7.4 then presents analysis of the annotated corpus and concluding thoughts.

7.1 ASPECTUAL VAGUENESS AND AMBIGUITY

As discussed in section 2.4.2, lexical aspect is projected from the verb; we have also seen in section 2.3.7 that verbs can be polysemous, having multiple word senses. It is eminently possible for these different word senses to belong to different lexical aspect classes. Thus, we should not be surprised that the same verb can occur in different contexts with different lexical aspectual classes.

Siegel and McKeown (2000) recognised that verbs can be aspectually ambiguous on the type level (i. e., different instances of the same verb lemma in different contexts may receive different aspectual class labels). However, more recent work has also examined the phenomenon of verbs that are aspectually ambiguous in context (i. e., on the token level). Friedrich and Palmer (2014), for example, explicitly allow for the possibility that a verb can be ambiguous for stativity in a single context (their example is 'Your soul was **made** to be filled by God himself,' p 517).

Croft, Peskova and Regan (2016) present a proposal to extend the Richer Event Description (Ikuta et al., 2014), an annotated corpus of time, modality, and event coreference, with PropBank framesets used for semantic role labels. The new annotation scheme they advocate would label each verb instance with an aspectual class, similar to those listed in Vendler's classification. In conducting preliminary feasibility studies for this new resource, they discovered that verb instances exhibit systematic aspectual ambiguity, even in context. They suggest that verbs that are ambiguous in context are likely to be at most two-way ambiguous for aspectual class.

I shall have reason to come back to this line of thought in section 7.3, where we will also see how to treat verbs that are not easily classified for aspectual class. One example of this is:

(7.1) wenn der Kunde die Karte abtrennt.
when the client the card removes.
when the client removes the card.

This appears to be the effects of underspecification: The verb does not seem to specify whether the event it refers to is punctual or durative. Other cases are less clear:

(7.2) diese Firmen zeigen bessere Ergebnisse
these companies show better results
these companies show better results.

Here, the aspectual type of the verb phrase could be stative (the companies have a tendency to outperform their competitors), or the phrase could be an extended change of state (after a period of concerted effort, the companies have achieved performance that is higher than some unexpressed point of comparison).

Another form of regular aspectual ambiguity is neatly illustrated by the family of *degree achievements* (Kennedy and Levin, 2008). Degree achievements are verbs morphologically derived from scalar adjectives (e. g., 'lengthen', 'shorten', 'widen', 'dry'):

(7.3) The soup cooled for an hour. (unbounded)

(7.4) The soup cooled in an hour. (extended change)

Here, we see a systematic ambiguity that licenses both durative and time-span adverbials; the two contingencies arise from the question of whether the soup is becoming relatively 'cooler', an unbounded process that could in principle continue indefinitely, or if it is becoming absolutely 'cool', a (durative) process that terminates, for example, at a particular temperature.

7.2 ASPECTUAL INDICATORS

My first efforts toward modelling computational aspect focused on translating the linguistic indicators of Siegel and McKeown (2000) into German; for reference, these are described in section 4.3 and listed in table 4.1. As I have noted, these indicators have been used successfully in more recent work on English aspect (e. g., Friedrich and Palmer, 2014).

It is worth mentioning at the outset that several of Siegel and McKeown's indicators do not translate well to German. Past and present participles do not indicate a distinction between events and static situations, for instance[1]. The present tense as a signifier of a stative or a habitual also does not work in German; dynamic verbs may be used in the present tense to indicate current or near future action. The perfect tense in German does not entail that the verb denotes a culminated event, as it does in English.[2] It can also be used to express a stative. The progressive form in English, indicative of an extended event, is, to all intents and purposes, lacking in German (cf. section 2.4.1). The *Rheinische Verlaufsform* (cf. section 2.4.1) is not attested a single time in the whole of the parsed SdeWaC corpus. I have also not found the combination of the adverb *gerade* ('currently, directly') with the present tense to be a good predictor of progressive aspect. While this can pick out *Sie greift gerade zum Speer* 'She is reaching for the spear', it also gets *den Kopf gerade halten* 'keep your head straight'. The simple past and the present perfect tenses are today used interchangeably in German, and do not carry an aspectual distinction.

Siegel and McKeown's 'not or never' indicator is approximated by a set of tests for various kinds of propositional negation:

- Negation with *nicht* 'not': *Da kann man sich* ***nicht*** *sicher sein.* 'You can't be sure about that.'
- Negation by negative argument: *Es gibt* ***keine richtige Ethik***, ***nichts Festes***. 'There is no correct ethics, nothing set in stone'
- Negation by governing verb: *Um zu* ***verhindern***, *daß . . .* 'To prevent . . . '
- Negation by negative adverb: *Keinesfalls*, *keineswegs*, *nie*, *niemals*, *nirgends*, *nirgendwo* 'No way, never, nowhere'.

The 'duration in-PP' indicator is translated as a set of prepositional phrases that should combine well with unbounded events:

1 Rather, this distinction can sometimes be indicated with the choice of auxiliary verb; compare stative *Der Verein ist geschlagen* 'The (football) club is defeated' and dynamic *Der Verein wurde geschlagen* 'The club has been defeated'.

2 This might possibly be approximated by searching for verb instances in the past tense in the presence of *gerade eben* 'just now' as a marker for relevance to the present moment of utterance; however, I have not tried this.

- *ab* 'from': ***Ab 1. Januar*** *zahlen Kunden 1 Euro mehr.* 'From January 1st, customers will pay 1 euro more.'
- *bis* 'until': *das* ***bis 1941*** *unter sowjetischer Verwaltung blieb.* 'which remained under Soviet administration until 1941.'
- *für* 'for': *Zweifellos hat das* ***für die längste Zeit*** *so gegolten.* 'Undoubtedly, that has been the case for the longest time.'
- *seit* 'since': ***Seit einigen Jahren*** *arbeitet er nicht mehr.* 'He hasn't worked for a few years.'
- *von . . . an* 'from . . . on': ***Vom ersten Augenblick an*** *beeindruckte sie ihn.* 'From the first moment she impressed him.'
- *von . . . bis* 'from . . . until': *Das fand* ***vom 23. bis 26. März*** *in Zürich statt.* 'That took place from March 23rd to March 26th in Zurich.'

Prepositional phrases using *für*, *seit*, *bis*, and *ab* were restricted to only those taking 'temporal' arguments by ensuring that the lemmatised version of their prepositional argument was contained in the following list: *Augenblick* 'moment', *Sekunde* 'second', *Minute* 'minute', *Stunde* 'hour', *Tag* 'day', *Nacht* 'night', *Woche* 'week', *Monat* 'month', *Moment* 'moment', *Mittwoch* 'Wednesday', *Winter* 'winter', *Sommer* 'summer', *Herbst* 'autumn', *Januar* 'January', *Februar* 'February', *März* 'March', *April* 'April', *Mai* 'May', *Juni* 'June', *Juli* 'July', *August* 'August', *September* 'September', *Oktober* 'October', *November* 'November', *Dezember* 'December', *Saison* 'season', *Epoche* 'epoch', *Ära* 'era', *Periode* 'period', *Jahr* 'year', *Jahrzehnt* 'decade', *Jahrhundert* 'century', *Jahrtausend* 'millennium', *Zeit* 'time'.

Similarly, the 'duration for-PP' indicator is translated to a set of prepositional phrases that should predict bounded events:

- *innerhalb* or *innerhalb von* 'within': *Das wurde* ***innerhalb der letzten zehn Jahre*** *geschaffen.* 'That was created within the last ten years.'
- *in* + cardinal: *Er hat das Buch* ***in 5 Minuten*** *gelesen.* 'He read the book in 5 minutes.'
- *in* + *einig* or *wenig* 'in a few . . . ': *Das endet* ***in einigen Monaten****.* 'That ends in a few months.'

The 'temporal adverb' indicator was translated to a set of adverbials describing the temporal qualities of an event:

- Adjectival durative cardinal + *lang* 'long': *Er schlief* ***zwei Tage lang****.* 'He slept for two days.' This indicator predicts unbounded events or statives; the rest of the indicators in this list express iteration or habituality.

- Adverbial *alle paar* 'every few': *Nur **alle paar Tage** besteht Gelegenheit*. 'Opportunity only comes every few days.'
- Adverbial *alle* 'every' + cardinal: *Sie treffen sich **alle zwei Jahre**.* 'They meet every two years.'
- Adverbial *jede* 'each' + temporal: *Die Leser freuen sich **jeden Samstag**.* 'The readers are happy each Saturday.'
- 'Grid' adverbials: *jährlich* 'yearly', *monatlich* 'monthly', *regelmässig* 'regularly', *täglich* 'daily', *tagsüber* 'during the day'.
- 'Quantification' adverbials: *häufig* 'often', *immer wieder* 'again and again' (*Also ich muß **immer wieder** sagen* 'I keep saying . . . '), *mehrfach* 'multiple', *mehrmals* 'many times', *oft* 'often', *oftmals* 'oftentimes', *selten* 'seldom/rarely'.

Other types of adverbs were also captured:

- Agentive adverbials: *absichtlich* 'intentionally', *allmählich* 'gradually', *bewusst* 'deliberately', *extra* 'on purpose', *gern* 'gladly', *gerne* 'with pleasure', *gezielt* 'purposefully', *lieber* 'preferentially', *persönlich* 'personally'.
- Agentive conjunction *um zu* 'in order to': *Sisko holt Eddington zu Hilfe, **um diese aufzuspüren**.* 'Sisko calls Eddington to help track them down.'
- Other adverbials: *kontinuierlich* 'continuously', *langfristig* 'long term', *langsam* 'slowly'.

The 'no subject' and 'evaluation adverb' indicators of Siegel and McKeown were not translated.

Table 7.1 summarises the aspectual indicators derived here from Siegel and McKeown's work.

Manual error analysis revealed several indicators from this translation effort that were not successful in describing temporal features of VPs. In particular, these prepositions had particularly low signal to noise ratios: *an (Tag)* 'on the day', *gegen (Uhr/Tag/Monat/Jahr)* 'around (hour/day/month/year)', *in* + temp NP 'in *X* time', *infolge* 'owing to', *nach* 'after', *vor* 'before', *während* 'during', *zwischen* 'between', and *über* 'over'. The adverb *hiermit* 'hereby' would have been useful as an indicator of explicit performative sentences, but it was unfortunately not attested at all in SdeWaC. The adverbials *plötzlich* 'suddenly' and *jede* + temporal NP 'every *X* time units' also did not perform very well. *Innerhalb (von)* 'within' can indicate temporal location, and not duration as intended, as in *Die Wahl erfolgt innerhalb der ersten drei Monate* 'The election takes place within the first three months.' The same is true for *von* 'from': *die Anschläge vom 11. September* 'the 9/11 attacks'. Finally, *für* can modify a NP instead of a VP, which can confuse the

Bounded	Unbounded	Negation	Agentive	Punctual/Durative	Iterativity
pp innerhalb von	pp für	part nicht	adv bewusst	adv plötzlich	adv regelmässig
pp in + temp	pp von bis	adv nie	adv gefälligst	adv allmählich	adv kontinuierlich
pp in + card temp	pp von an	adv niemals	adv gern/e	adv kurz	adv jährlich
pp in + einig/wenig/kurz	pp seit	adv nirgends, nirgendwo	adv lieber	adv rasch	adv monatlich
	pp bis	adv keinesfalls, keineswegs	adv ungern	adv langsam	adv täglich
	pp ab	adv beinahe, fast, nahezu	adv persönlich	adv langfristig	adv mehrfach
	adjp temp + lang	subj keine/r/s	adv gezielt		adv mehrmals
		subj niemand	adv extra		adv abends
		obj nichts/nix	adv absichtlich		adv morgens
		obj kein/e/en/s	conj um zu		adv nachts
		obj niemanden			adv tagsüber
		gov verhindern			adv häufig
		gov untersagen			adv oft/öfters/oftmals
		gov verweigern			adv selten
		gov verbieten			adv immer wieder
		gov ablehnen			advp alle paar + temp
					advp alle card + temp
					advp jeder + temp

Table 7.1: Summary of aspectual indicators for German.

automatic parser and lead to errors in syntactic analysis: *einen Antrag für die Jahre ab 1983* 'an application for the years from 1983 onward'.

I also added several extensions to Siegel and McKeown's scheme. The first was a series of indicators for various kinds of plural: indicators indicating plural subject, plural object, dative, PP argument, or plural verb conjugation. I will use these as a quick way to filter out bare plurals, which can have effects on the aspectual class of a verb phrase (cf. section 2.4.7 and appendix A). The second set of new indicators was based off of Herweg (1991), who presents a mereological calculus of periods of time in German based on relations of inclusion and precedence. This article contains a list of conjunctions that can have a temporal interpretation, of which I use several here: *bevor* 'before', *bis (zu)* 'to (up to)', *indem* 'whilst', *seitdem* 'since', *solange* 'as long as', *während* 'while'. The conjunctions *als* 'as, when', *nachdem* 'after', *sobald* 'as soon as', *wenn* 'when' were not found to be helpful. In the way I have done things here, each conjunction produces two aspectual indicators, one for the main clause (with the suffix `_top`), and one for the subordinate clause (`_bottom`).

7.2.1 *What do the indicators indicate?*

Indicators were operationalised by defining search criteria on the structure of the parsed sentence, making use of the lexical and morphological analysis delivered by the parser. I ran all aspectual indicators on each verb instance in the SCF lexicon developed in section 5.3. This produces a large database associating each verb instance in the corpus with a list (possibly empty) of aspectual indicators that are active for that instance.

What verbs do particular kinds of indicators like? Table 7.2 shows the most common verbs found with each kind of aspectual indicator, where indicators are grouped together (summed up) in the way that they are organised in table 7.1. In an attempt to avoid any spurious effects from aspectual transformation or coercion, all verb instances that occur with a plural (subject, object, etc.) or in a negated context are excluded from the count. The indicators selecting for boundedness seem to be of reasonable quality: The VPs *sich entwickeln zu* 'develop into', *führt zu* 'leads to', and *(es) kommt zu* 'it comes to' can all plausibly be telic bounded phrases. Similarly, the unbounded indicators appear to work well enough: *Stattfinden in/von* 'take place in/from', *leben* 'live', and *arbeiten* 'work' are stative or unbounded. The quality of the indicators is less clear with the agentive grouping, where the verbs selected are quite general in meaning, and with the iterative/habitual grouping, which is also difficult to characterise. Note that *geben* `xa` (existential 'there is') co-occurs frequently with many kinds of indicator; this is a reminder of the frequency of this construction in German, as well as of the semantic flexibility of the expression.

Verb	SCF	Count
Bounded indicators: pp_innerhalb, pp_in_card, pp_in_einig.		
entwickeln	npr:zu.Dat	140
geben	xa	136
führen	np:zu.Dat	100
kommen	np:zu.Dat	66
kommen	xp:zu.Dat	52
Unbounded indicators: pp_fuer, pp_vonbis, pp_vonan, pp_seit, pp_bis, pp_ab, ap_lang.		
geben	xa	2268
stattfinden	np:in.Dat	2092
stattfinden	np:von.Dat	1658
leben	np:in.Dat	1300
arbeiten	np:in.Dat	686
Agentive indicators: adv_absichtlich, adv_bewusst, adv_extra, adv_gern, adv_gerne, adv_lieber, adv_persoenlich, adv_gezielt.		
machen	na	774
sehen	na	627
tun	na	486
haben	na	343
werden	ni	337
Iterative/habitual indicators: adv_allmaehlich, adv_ploetzlich, adv_langsam, adv_langfristig, adv_haeufig, adv_immerwieder, adv_jaehrlich, adv_kontinuierlich, adv_mehrfach, adv_mehrmals, adv_monatlich, adv_oft, adv_oftmals, adv_regelmaessig, adv_selten, adv_taeglich, adv_tagsueber, advp_alle_card, advp_alle_paar, advp_jeder_np.		
vorkommen	n	1701
geben	xa	1162
handeln	xr	920
kommen	n	725
sehen	na	686

Table 7.2: Top verb-SCF combinations selected by each group of aspectual indicators.

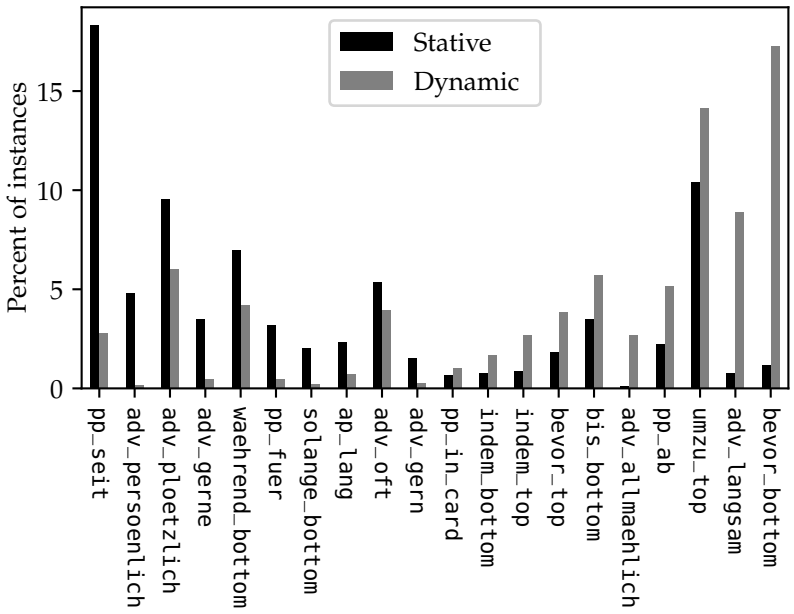


Figure 7.1: Linguistic indicators that are particularly stative or particularly dynamic.

Which indicators are activated by particular kinds of verbs? Here, I will focus on a single aspectual category, namely the distinction between stative and dynamic events. As a rough approximation, I take a small sample of verbs to represent dynamic events: *beginnen* ('begin', punctual change), *drehen* ('turn', unbounded or extended change), *merken* ('notice', punctual change), and *zusammenfinden* ('come together', extended change). Stative verbs are similarly represented by a small verb sample: *kennen* ('know', stative), *glauben* ('believe', stative), and *stehen* ('stand', a semistative). As before, I exclude verbs with plural arguments, or those that occur in negative contexts.

I sum up the indicators that are active for the given verb types, segregating the counts by aspectual category, and L1-normalise the resulting observations to obtain percentages. Thus, for example, 18% of the time, when an indicator co-occurs with the stative verbs, the indicator turns out to be `pp_seit`; in contrast, this happens only 3% of the time when the indicator co-occurs with one of the dynamic verbs.

Figure 7.1 shows the aspectual indicators, ordered by the difference in co-occurrence rates between the stative and dynamic categories. On the left are those indicators that are more strongly associated with statives, and on the right those that rather prefer dynamic verb instances. Only the extremes of the scale are shown; those indicators in the middle of the ranking that have no strong preference for one category or the other are left out.

The figure shows several behaviours that agree well with theoretical predictions:

- `bevor_bottom` means the verb occurs in a subordinate clause headed by the conjunction *bevor* 'before': The temporal threshold when the verb of the main clause comes to an end is likely to be introduced by an event rather than a state. The same holds for `bis_bottom`.
- `umzu_top` means the verb occurs in a main clause that includes a subordinate clause linked with *um zu* 'in order to'; hypothetically, this main clause verb should be agentive (and hence dynamic).
- PP *in* + cardinal + temporal NP should be associated with boundedness, and, indeed, it occurs more often with dynamic verbs than with stative ones.
- Temporal adverbs *langsam* and *allmählich* pattern with dynamic verbs, as expected.
- PPs that are incompatible with bounded events may be more likely to pattern with statives, and we see this to some degree: PP-*seit*, PP-*für*, temporal NP + *lang*.

There are also some findings here that are unexpected:

- *plötzlich* should not happen with statives; and,
- *gern* (agentive) and *oft* (iteration) should be more likely to happen with dynamic verbs than stative ones.

Finally, the indicators *seit* and *persönlich* are likely influenced in this study by lexical effects of the verb *kennen* 'know' ('I know him since'; 'I know him personally'), and do not represent temporal adjuncts that indicate aspectual behaviour.

7.2.2 *Clustering indicators*

Figure 7.2 shows the results of a brief study to look at the coherence of the aspectual indicators, namely to what degree different indicators in the same grouping exhibit similar patterns. This closely follows the clustering procedure I have used in chapters 5 and 6. Here I follow the intuitive hypothesis that bounded prepositional phrase indicators should pattern similarly to each other; and the same should be expected from the unbounded prepositional phrase indicators. Taking the co-occurrence counts from before (i. e., those shown in table 7.2), I represent each aspectual indicator as a vector whose dimensions are the various possible verb-SCF types. I normalise these vectors so that they represent discrete probability distributions, and then perform hierarchical clustering using Ward's criterion, with the Jensen-Shannon divergence as the distance measure.

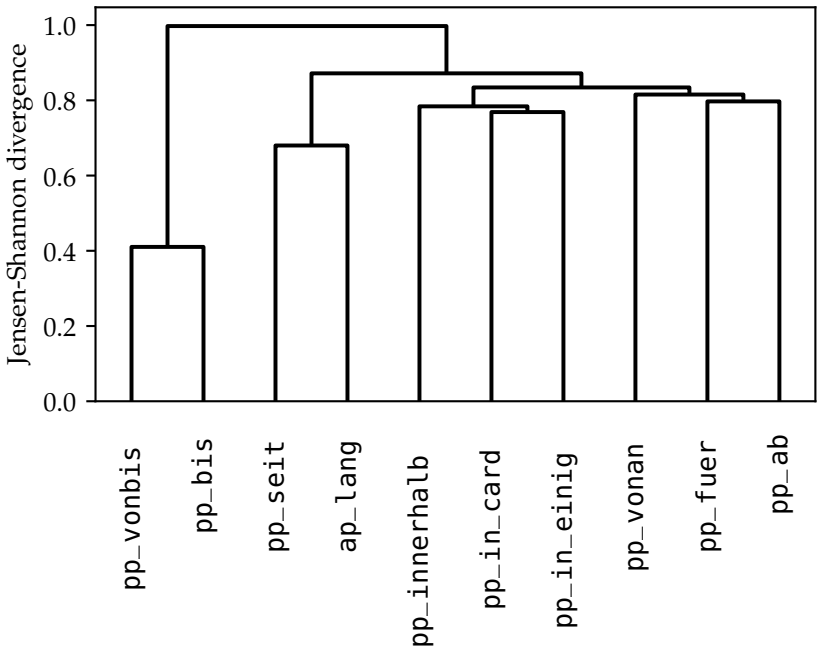


Figure 7.2: Hierarchical clustering of bounded and unbounded PP indicators.

The dendrogram displayed shows just the indicators selecting for bounded or unbounded events. We can see that the three indicators that should pick up bounded events (*innerhalb*, *in* + cardinal, and *in einige* + temporal NP) are indeed grouped together by this analysis; however, the indicators for unbounded events are much less clearly grouped. *Von … bis* and *bis* are close together; as are PP-*seit* and NP *lang*; and *von … an*, *für*, and *ab* make up a third family. These subgroupings appear to be quite dissimilar from each other. This raises the question of whether the aspectual indicators developed here can be considered reliable: If we cannot use these indicators to discern a single aspectual feature of verbs, then they may be of little use to us.

To better understand these findings, one can compare which verbs commonly co-occur with the different indicator groupings. The *von … bis* and *bis* indicators happen with *stattfinden* 3769 times ('to take place'), *dauern* `np:von.Dat` 247 'to last', and *laufen* `np:von.Dat` 150 'to run'; these are all verbs expressing unbounded events. The PP-*seit* and NP *lang* grouping has: *geben* `xa` 1765 'there is', *leben* `np:in.Dat` 1239 'live', *versuchen* `ni` 647 'try to', *arbeiten* `np:in.Dat` 625 'work', *bestehen* `n` 578 'exist'. These verbs are a mix of unbounded and stative. Finally, *von … an*, *für*, and *ab* have: *gelten* 779 'to be valid', *geben* `xa` 400, *rechnen* `np:für.Acc` 302 'calculate', and *erwarten* `nap:für.Acc` 218 'expect'. Here we see in the case of *rechnen* a verbal argument incorrectly detected as a durative adverbial adjunct. It is also notable that there is little overlap between the verbs selected for by these

indicators; this may be due to the low counts of verbs found by each indicator.

7.2.3 *Summary*

Ultimately, the indicator approach is defeated by the low frequencies of the indicators in the SdeWaC corpus.

The most common indicators in SdeWaC are *um zu* with 218,351 occurrences, PP-seit with 84,773, and adverbial *oft* at 57,542. Given that there are 82,873,358 verb instances total in SdeWaC, this represents incidence rates of 1 in 380, 1 in 978, and 1 in 1440 verb instances, respectively. The least common indicators are *immer wieder* with 393 instances (1 in 210,000), and *alle paar* with only 64 (1 in 1.3 million verb instances).

The low co-occurrence counts restrict me to analyses of verbs on the type level, as in section 7.2.1; there are simply not enough verb instances to make statistical predictions on the token level. Thus, this makes it hard to investigate the role of verb sense, aspectual ambiguity, or aspectual coercion.

Considering that section 7.2.2 also raises questions about the reliability of the aspectual indicators, it seems that it is time for a new strategy. Section 7.3 will describe work to create a manually annotated corpus of German verb instances to better pin down the category of aspect.

7.3 ANNOTATED CORPUS

Apart from the sorts of linguistic indicators we have seen in the previous section, lexical aspect is a category that does not manifest syntactically. Unfortunately, there does not exist a prior linguistic resource on lexical aspect in German. Therefore, I set out in this section to produce one.

The corpus I will describe here consists of a set of verb instances in their clausal contexts, manually annotated for features of lexical aspect. These indicate, for example, if the situation referred to is bounded or unbounded. The corpus is annotated on the token level, by considering individual verb instances; such a token-based resource can be collapsed to a type-based resource by grouping observations by which verb they occur with. As with the construction of any corpus, care and consideration are crucial when curating its contents; it is desirable for the verbs in the corpus to be balanced for frequency and, ideally, also for aspectual class.

Each clause in the corpus contains a main verb to be annotated for its aspectual class. Verb instances were chosen from SdeWaC such that:

- no clause to be annotated had more than 50 words;
- no verb to be annotated co-occurred with the negative particle *nicht* 'not';
- no verb to be annotated was a modal or auxiliary verb, and no verb to be annotated was governed by a modal verb; and,
- every verb chosen was required to be a real word with correct spelling, by requiring that it appear in GermaNet (section 3.3).

No other filtering was performed. In particular, this means that the the corpus attests several multi-word expressions, representing idioms and light verb constructions; examples are *im Blick haben* 'to have an eye on', *in Anspruch nehmen* 'to take advantage of', *Besuch bekommen* 'get a visit', and *zum Opfer fallen* 'fall victim to'.

The following section 7.3.1 describes the features annotated on these verb instances. After that, section 7.3.2 details how the verb instances were chosen to become part of the corpus. Section 7.3.3 describes the annotation tool, and section 7.3.4 contains some notes on the annotation process. Section 7.3.5 presents a study to measure how well individual annotators agreed with each other. Section 7.3.6 gives some summary statistics and analysis of the completed corpus. Section 7.3.7 reconsiders the topic of aspectual ambiguity and proposes a solution in the form of 'ambiguity classes'. Finally, section 7.3.8 examines whether aspectual ambiguity correlates with polysemy, and takes a look at verbs that are aspectually unambiguous.

7.3.1 *Aspectual classes*

The annotation project I am describing here would not go very far without a design for the aspectual information that annotators should provide about the corpus. For this I will propose a system of classification of aspectual classes, developed together with my supervisor, Prof. Dr. Markus Egg. This classification is a combination of the aspectual type inventories of Vendler (1957), Moens and Steedman (1988), and Egg (2005).

Following the description of aspectual categories I have outlined in section 2.4, situations described by verbs are first divided into states and dynamic events. Dynamic events are then further classified as to whether they are unbounded (like Vendler's activities) or bounded (having a built-in ending, as well as an intrinsic expected run time). Bounded events can be classified on two dimensions. Firstly, they can require that a change of state happens in the world, or not. And secondly, they can be punctual or extended in time. This categorisation procedure, depicted in figure 7.3, yields six classes:

- STATES;

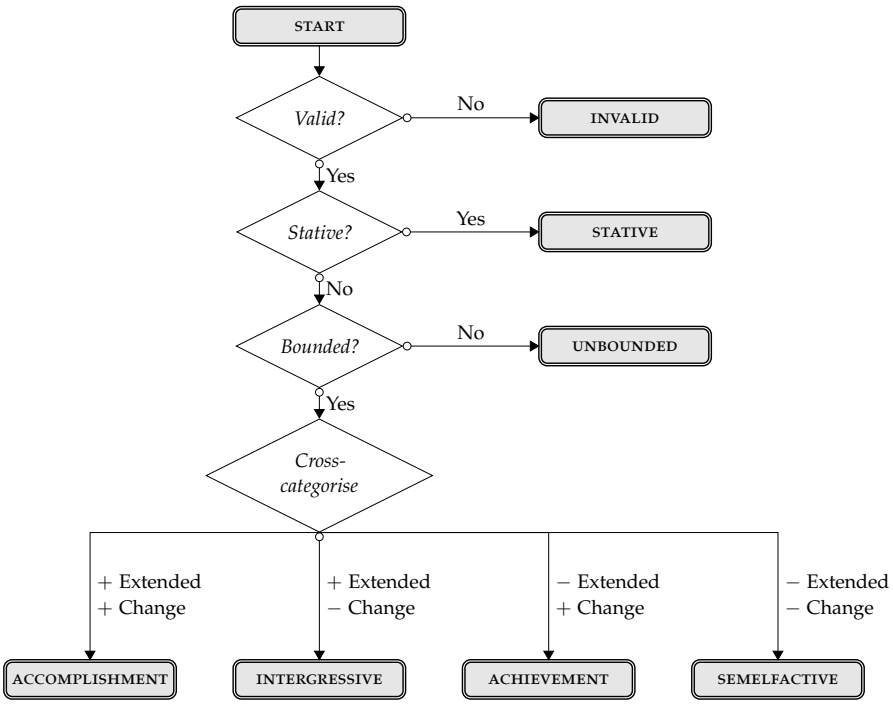


Figure 7.3: A taxonomy of aspectual classes.

- UNBOUNDED events;
- ACCOMPLISHMENTS: extended events with a change of state;
- INTERGRESSIVES, following Egg (1995): extended events with no change of state;
- ACHIEVEMENTS: punctual events with a change of state; and
- SEMELFACTIVES, following Smith (1991): punctual events with no change of state.

The aspectual classification of Moens and Steedman (1988) is very much like the six classes I use here, but is missing the distinction between *unbounded* events ('to gaze at the sunset') and *intergressives* (extended no change events, 'to watch a film').

The category of INVALID is used to filter out cases resulting from parser errors where the main verb to be annotated is an auxiliary or modal, or in cases where the wrong verb is identified (e. g., missing verb prefix). Also invalid is any clause that makes no sense or cannot be interpreted in some way, as can happen with clause segmentation failures, which can cause words to be missing.

Table 7.3 shows how this aspectual classification is related to the aspectual type inventories of Vendler (1957), Bach (1981), Moens and Steedman (1988, M&S), and Egg (2005). This taxonomic compatibility will turn out to be a very useful feature of the classification, as we shall see.

	stative	dynamic				
		unbounded	bounded			
			ext./no-ch.	ext./ch.	punc./no-ch.	punc./ch.
Vendler	state	activity	accomplishment		achievement	
Bach	state	process	event			
M&S	state	process		culminated process	point	culmination
Egg	stative	process	intergressive	change	intergressive	change
This thesis	state	unbounded	intergressive	accomplishment	semelfactive	achievement

Table 7.3: Mapping of aspectual classes to previous classifications from the literature.

Hypotheses are best advanced before the work is begun in earnest. To this end, it is good to recall at this point the argument from Egg (1995) that the quality of being extended or punctual should be viewed as gradual and not binary, unlike the other feature dichotomies. This prediction licenses the hypothesis that this category should be more difficult to annotate than the others, because the distinction between what is a 'momentary' predicate and what is not may be fuzzy.

7.3.2 *Corpus composition*

The corpus was created from three parts. Part A (3000 clauses) used a verb sample; part B (900 clauses) did not use a verb sample; and part C (300 clauses) was specifically focused on clauses containing intergressive (durative no-change-of-state) and semelfactive (punctual no-change-of-state) verbs. Parts A and B are therefore broadly representative of verb types found in the SdeWaC corpus, but Part C is decidedly not.

The following sections describe these parts in more detail.

7.3.2.1 *Part A*

Part A of the corpus was constructed using a verb sample; 60 verb types were chosen as representatives, and multiple instances of these were collected to sample their possible senses and uses.

The 60 verbs were chosen to include:

- 20 high-frequency verbs (randomly drawn from the 65 verb types with counts $> 10^5$ in SdeWaC);
- 20 medium-frequency verbs (drawn from the 602 verbs with counts $> 10^4$); and
- 20 low-frequency verbs (drawn from the ~2100 verbs with counts $> 10^3$).

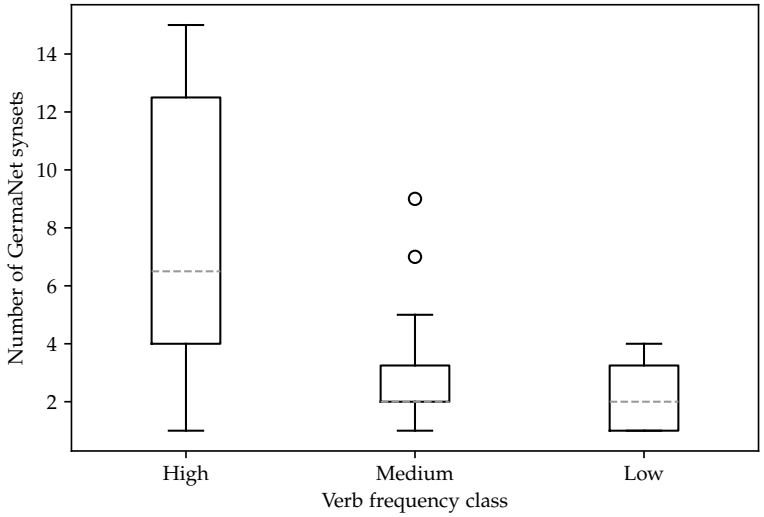


Figure 7.4: Number of GermaNet synsets by verb frequency class for the aspectual corpus verb sample (corpus part A).

For each verb in this sample, 50 clauses containing that verb were chosen at random from SdeWaC.

The sample of verbs used in part A of the corpus present an opportunity to study the interaction of verb frequency and polysemy. For each of the 60 verbs used in part A, I record how many synsets in GermaNet include that verb; this value is then used as a proxy for the polysemy of the verb. Figure 7.4 shows the number of verb senses as a function of the verb's frequency class, confirming that higher-frequency verbs are positively correlated for polysemy, as we would expect. This effect also holds more generally of all the verbs in SdeWaC, and this finding supports the conclusion that the verb sample used in part A is representative of German verbs in general, and is balanced for verb polysemy. Note that the high frequency group still contains one verb with only a single synset (*entstehen* 'arise'), so that the annotated corpus can be used to check how an unambiguous high frequency verb behaves.

7.3.2.2 *Part B*

Part B did not use a verb sample. Instead, I randomly selected 300 clauses containing a high frequency verb, 300 sentences with a medium frequency verb, and 300 sentences with a low-frequency verb. There was no constraint on this choice other than that the verb did not appear in the verb sample used to create part A. While the clauses in part B are balanced for frequency, they do not represent a lexical sample like the verbs in part A; rather, they can be seen as a balanced corpus sample, in some ways more representative of the contents of

the SdeWaC corpus. The result is that part B contains 519 verb types, whose distribution is Zipfian (cf. figure 7.5): Most verbs in part B occur only a few times, and a very few verbs occur many times. The most common verb in part B is *geben*, because the existential construction *es gibt* 'there is' is very common in German.

7.3.2.3 *Part C*

Part C of the corpus was designed to capture more instances of intergressive verbs (extended no change) and semelfactives (punctual no change). Parts A and B, although balanced and representative of actual language use, systematically underrepresented these categories, because they appear to be naturally rare compared to the other aspectual classes.

Because determining the aspectual class of a given verb is nontrivial, collecting random instances of intergressives and semelfactives was a challenge. I followed two approaches here:

1. Produce a list of verbs that tend to be intergressives and/or semelfactives (a verb sample), and collect sentences containing these verbs; and

2. using the data from parts A and B that had already been manually labelled, train a supervised classifier to detect intergressive and/or semelfactive verbs, and use this classifier to sample verbs from SdeWaC that the classifier judged most likely to be from these categories.

The second option was used to collect intergressive verbs for part C, but this was unsuccessful for semelfactives. Using annotations already completed on parts A and B, a version of the maximum entropy aspectual classifier described below in section 8.1 was trained, including word embeddings (`sdewac-lemmas`, cf. section 8.1.8) as features. The classifier achieved a micro F1 score of 86.7% under 3-fold cross-validation on parts A and B of the corpus. The classifier was then used to automatically label 10K unseen clauses for aspectual class drawn at random from SdeWaC. The 150 verb instances that the classifier was most confident were intergressives were then included in part C; as will be made clear by figure 7.7, this enterprise was largely successful. Note that this procedure will necessarily tend to choose clauses that are similar to those already in parts A and B, and thus may not be very good at increasing the linguistic diversity of the corpus; however, this method is neutral, impersonal, and impartial, and does not rely on potentially flawed human intuitions of what constitute good intergressives.

At the time of the construction of part C, there were only 10 clauses in parts A and B that had been manually judged to be semelfactives; of the 10K random clauses drawn from SdeWaC, the classifier

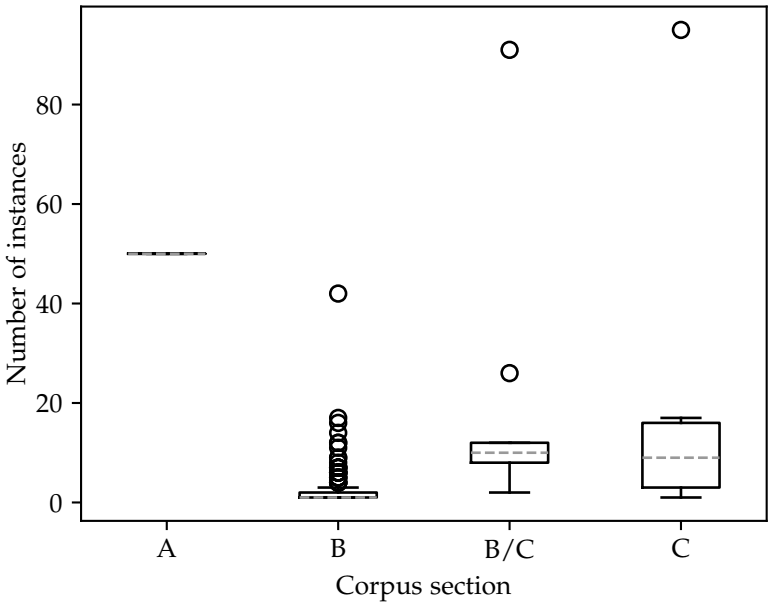


Figure 7.5: Number of instances of each verb in the annotated corpus, broken down by section.

only picked out two clauses that it believed to be semelfactive, and a manual examination of these did not look promising.

Therefore, I manually created a sample of six verbs to collect the semelfactive verb instances: *niesen* 'sneeze', *blinken* 'flash', *aufblitzen* 'flare up', *husten* 'cough', *klopfen* 'knock', *blinzeln* 'blink'. All of these verbs, save *klopfen*, were not frequent enough to be included in the 'Low Frequency' class of verbs, but were found in a new frequency class, the 'Very Low Frequency' verbs. *Niesen* was not attested at all in SdeWaC. Note that a collection of semelfactives constructed using a verb sample may be incomplete because the verb sample may overlook particular usages or contexts.

As a result of this work, part C is a collection of 300 clauses documenting two rare aspectual classes, and containing 18 German verb types. 150 clauses contain semelfactive (punctual no-change verbs), with a total of five verb types. Another 150 clauses contain intergressive (extended no-change) verbs, with a total of 13 verb types. Nine of these 13 overlap with verb types already represented in part B (but none of the verb instances in part C are duplicated from part B). This is evidence that the classifier did in fact pick verbs that seemed familiar from its training data.

Figure 7.5 shows the distribution of the verb types included in the corpus, broken down by which part of the corpus the verb is found in. The box-plot labelled 'A' shows the distribution of the 60 verb types that are found *only* in part A of the corpus, and box-plots 'B' and 'C' do the same thing for their respective parts (510 and 9 verbs,

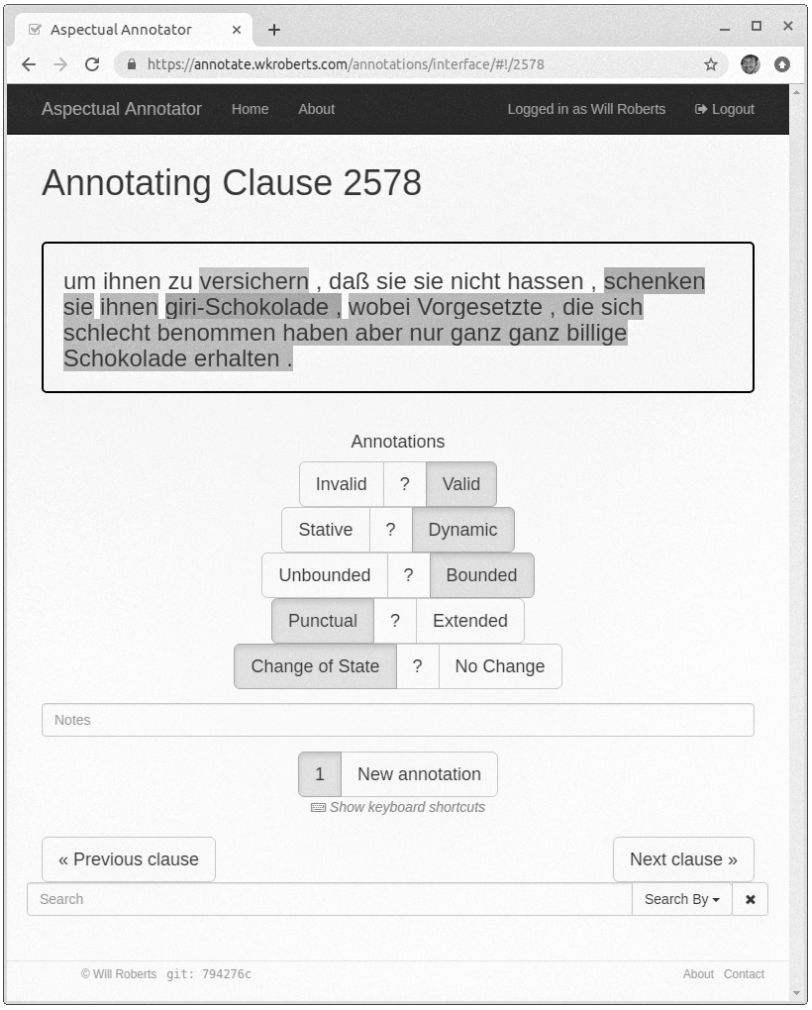


Figure 7.6: Aspectual annotation web application.

respectively). The plot labelled 'B/C' shows the nine verbs that are common to both parts B and C.

The figure shows the difference between a verb sample (part A, with a fixed number of verb instances per type) and a random sample (part B, with a Zipfian distribution over verb types). The distribution of the verbs found only in part C is more like a verb sample than part B is, because most of these verbs are the semelfactives that were in fact constructed using a verb sample. The verbs shared between parts B and C are dominated by verb instances from part C, but distributionally they look much like the verbs in part B, with more skewed counts like one would expect from a random sample.

7.3.3 *Annotation tool*

Aspectual annotation was performed using a web application, depicted in figure 7.6. The application uses colour coding and mouseover

hints to show sentence structure and highlight potentially useful information. In the figure, *schenken* 'give' is coloured red, indicating that it is the main verb of the sentence, the thing that is to be annotated. The direct object *giri-Schokolade* 'giri chocolate' is drawn in green, and the subject *sie* 'they' and indirect object *ihnen* 'them' are drawn in blue, as is the adjunct PP beginning *wobei Vorgesetzte* 'although the management'. The aspectually relevant subordinate clause starting *um zu versichern* 'in order to ensure' is shown in yellow.

The interface, a single page application written in Javascript with the Angular framework, is designed to be as simple as possible. The number of the clause being currently annotated is displayed prominently on the page, and is also indicated in the location bar, allowing the user to quickly jump to a desired clause. The possible actions the user can take are indicated with a set of buttons. The interface enforces valid combinations of the aspectual categories represented by these buttons, following the taxonomy shown in figure 7.3. For instance, marking a clause as stative will automatically set the values of the other buttons to valid, unbounded, extended and no change, and disable the unbounded, extended and no change buttons so that their values cannot be altered. Buttons at the bottom of the page allow the user to jump back and forth between adjacent clauses.

The tool's primary design focused on speed of annotation. It made use of a mobile-first interface, allowing the app to be used on smartphones and tablets, permitting annotation to be more easily carried out during short breaks, while taking the subway, etc. The clause to be annotated is displayed in large font, and colours rapidly present important information. Most functions can be quickly carried out using single-letter keyboard shortcuts:

ESC Reset annotations

LEFT ARROW Previous clause

RIGHT ARROW Next clause

I Mark as invalid

V Mark as valid

S Mark as stative

D Mark as dynamic

U Mark as unbounded

B Mark as bounded

E Mark as extended

P Mark as punctual

C Mark as change of state

N Mark as no change

The tool's second design criterion was reliability. On the server, written in Python with the Flask framework, data were stored in an append-only format, so that an action taken by a human annotator would never be overwritten or deleted; rather, the annotated state of a sentence could be updated by the human at a later time, but the full log of annotation actions taken by the user was always available. This also makes possible various analyses of how often a particular annotator changed their minds, what times of day they were most productive, and how the reliability of their annotations changed over time.

Apart from enforcing the internal logic of the aspectual type taxonomy, the tool purposefully did not constrain the rest of the annotation. Categories can be annotated as 'Yes', 'No', or 'Unsure' (the latter representing not applicable, undecidable, under-specified, etc.). The annotator also has the ability to give a single sentence multiple annotations; this option was useful for cases of regular aspectual ambiguity. Multiple annotations are created simply by adding a new annotation on a clause. There is no method for deleting annotations; rather, the user is required to set the two annotations to be identical, which then causes the redundant annotation to be filtered out.

Finally, there is a facility for annotators to make notes to themselves, to facilitate future review; and a rudimentary search interface allows users to visit only clauses containing a particular verb, or containing a particular word.

7.3.4 *Annotation process*

The 4,200 clauses of the corpus were annotated by two annotators: one was a university graduate and native speaker of German, trained on the annotation task; and the other was the author of this thesis, a fluent second-language speaker of German.

Annotators were instructed to give the fundamental aspectual class of the verb as it appeared in context. Here the fundamental class follows Siegel and McKeown (2000) and means the aspectual type of the verb before any aspectual coercion and ignoring the influence of aspectual operators. A verb instance is classified according to five features (validity, stativity, boundedness, durativity, and change of state); this mirrors the hierarchy depicted in figure 7.3.

As mentioned in section 7.3.3, the aspectual features may be underspecified. A single verb instance may also receive more than one aspectual annotation, reflecting aspectual ambiguity. This is common with degree predicates, like the following:

(7.5) (unbounded or accomplishment) *trocknet die Haut zusätzlich aus.* 'also dries out the skin.'

The annotation took into consideration the direct object or patient of the verb if necessary. For example, a punctual change of state verb with a plural patient would be marked as extended if it is judged to be implausible that all patients were changed in the same instant:

(7.6) (accomplishment) *während sein Komplize mit einem Hammer zwei Ausstellungsvitrinen* ***einschlug****, daraus etwa 20 Uhren der Marke "Cartier" nahm und in einer schwarz-grün-blauen Sporttasche verstaute.* 'while his accomplice **smashed in** two display cases with a hammer, took about 20 "Cartier" watches from them and stowed them in a black, green and blue sports bag.'

Communication verbs denoting speech acts are not considered to lexically require that their perlocutionary act obtains. For example, in

(7.7) (intergressive) *Er ruft Scully zu sich.* 'He calls Scully over.'

the mere act of calling someone can imply but does not necessarily entail that they hear, respond, obey, or move from their current location; thus, the verb instance is marked as not having a change of state.

Textual or fictional agents can produce stative verb instances, because the verb is describing one or more qualities of a created work:

(7.8) (stative) *weil in den Werken Kandinskys immer wieder Reiter auftauchten.* 'because riders kept appearing in Kandinsky's works.'

Annotators were also instructed to attend to metaphorical usages; in such cases, the aspectual class should be chosen which most closely matches the intended meaning. Anecdotally, metaphorical uses of verbs seem to preserve the lexical aspect class of the literal verb.

Guidelines for the annotation were developed during the annotation process, and are reproduced in appendix A. These explain the aspectual features, and list some tests for how to determine the correct annotation of a verb. For example, the guidelines define the notion of Incremental Theme (cf. section 2.4.7), and point out that adverbials expressing volition are incompatible with stative predicates:

(7.9) ? Virginia eagerly/reluctantly/deliberately knew Haitian Creole.

All disagreements between the two annotators were adjudicated.

Both annotators took very close to 29 hours in total to annotate all 4,200 clauses between October 2017 and February 2019; they were also both very close to an average of 20 seconds per annotation.

Also, both annotators are very close in terms of how often they change their minds, with both changing about 25% of their annotations over time (via reconsideration and/or the adjudication process).

Invalid	$\kappa = 0.190$
Stative	0.746
Bounded	0.735
Change of state	0.758
Extended	0.292
Class	0.548
Class w/o extended	0.651

Table 7.4: Inter-annotator agreement on the aspectual class annotation.

7.3.5 *Inter-annotator agreement*

The agreement between the two annotators was measured after both annotators had annotated ca. 2,200 clauses and adjudicated their differences; this long training period, representing about the midway point of the annotation effort, should have given the annotators adequate opportunity to become proficient in the labelling task. For this measurement, both annotators annotated 248 unseen clauses. Table 7.4 shows the Cohen's κ between the judgements made by the two annotators. The first five rows show agreement on the five categories that are directly annotated using the tool; for example, the 'Bounded' row shows inter-annotator agreement about whether a given verb instance was bounded or unbounded. Nine clauses were marked as invalid by one or the other annotator, and these nine clauses were removed; all rows in the table after the first are calculated on the remaining 239 clauses. The 'Class' row shows inter-annotator agreement on the original six-way aspectual classification; and the 'Class w/o extended' row shows agreement when dropping the problematic punctual/extended feature (giving a 4-way classification).

Using the terminology of Landis and Koch (1977) to summarise these results, annotator agreement is substantial for the categories Stative, Bounded, and Change of State, but only fair for the category Extended. Agreement on the 6-way 'Class' classification is moderate, rising to substantial on the 4-way 'Class w/o extended' classification. Note that inter-annotator agreement on the Stative category is very similar to that reported by Friedrich and Palmer (2014), who had two annotators mark 6,161 English verb clauses for stativity, and found $\kappa = 0.7$; on a second corpus of 2,667 clauses containing aspectually ambiguous verbs they found $\kappa = 0.6$. Similarly, the inter-annotator agreement on the Change category is comparable to the $\kappa = 0.8$ reported by Siegel and McKeown (2000) when constructing their dataset annotated for culmination.

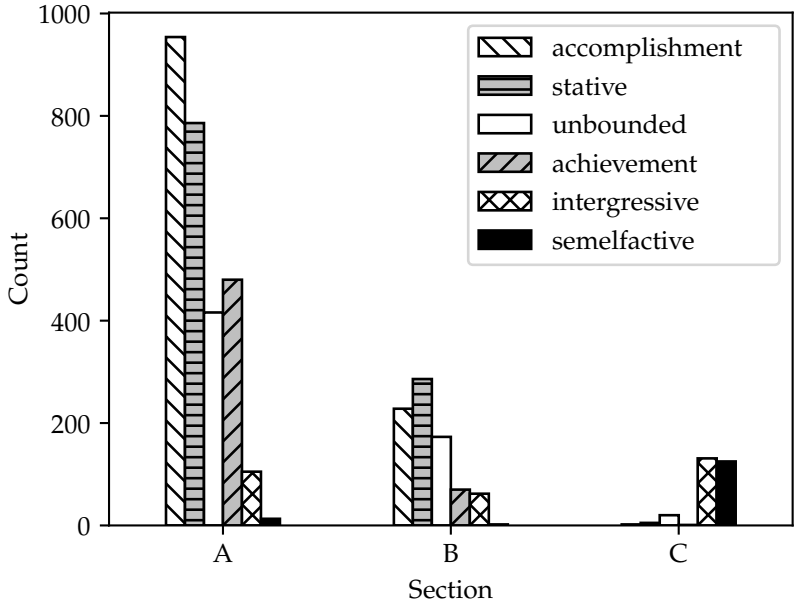


Figure 7.7: Number of instances annotated for each aspectual class in the annotated corpus, broken down by section.

7.3.6 *Corpus statistics*

The finished annotated corpus contains 4,052 valid verb instances, and 148 that were judged to be invalid. There are 3,060 clauses containing finite verb instances, and 1,140 with non-finite verbs. 3,567 verb instances are in the active voice, and 633 are passive. Of the 4,052 valid verb instances, 193 instances receive an annotation with either underspecified labels or multiple aspectual classes. 3,933 valid verb instances received only a single annotation; 117 instances received two; and 2 instances were given three annotations each. 1,380 verb instances fall in the high verb frequency class; 1,367 are of medium frequency; 1,398 instances are low frequency; and 55 are very low (i. e., in part C).

Figure 7.7 shows a summary of the results of the annotation, broken down by corpus part. For the sake of clarity, the figure does not include the approximately 8% of annotated clauses that did not receive one of the six most common labels (for example, about 3.5% of clauses were marked as invalid, and are not shown here). Comparing part A (using a verb sample) to part B (using a random sample), it is apparent that part A contains more accomplishments and achievements. We can also see that neither part A nor part B has very many instances of intergressives and semelfactives. This was the motivating factor behind creating part C; and, indeed, we can see that part C does end up being largely composed of intergressives and semelfactives. In fact, of the 150 clauses in part C that were chosen to likely contain semel-

Ambiguity class	Number of verb instances
accompl	1184
stative	1077
unbounded	609
achieve	551
intergressive	298
invalid	148
semelfact	140
accompl/achieve	59
accompl/unbounded	51
accompl/stative	31
intergressive/stative	12
intergressive/semelfact	9
achieve/stative	9
stative/unbounded	6
is???	5
achieve/unbounded	4
intergressive/unbounded	3
achieve/semelfact	2
achieve/stative/unbounded	1
accompl/stative/unbounded	1

Table 7.5: Ambiguity class frequencies resulting from the annotation effort.

factive verbs, 123 were judged to be actually semelfactive; similarly, of the 150 clauses chosen to capture intergressives, 129 were judged to be in that category.

7.3.7 *Ambiguity classes*

Underspecification and multiple annotation can be seen as complementary descriptions of the same phenomenon. For example, a common case is a change of state verb that could plausibly be either punctual or durative:

(7.10) (accompl/achieve) *Klink hatte das ausgelöste Filet mit Meersalz gewürzt und stramm in Klarsichtfolie eingerollt.* 'Klink had seasoned the separated fillet with sea salt and rolled it tightly in cling film.'

Here, seasoning a cut of meat is perhaps intended to mean a very brief application of salt, but it is also conceivable that the seasoning was done slowly and carefully over an appreciable time span. As a result, one annotator might attest that the verb instance is underspecified for duration. The other might argue that the verb instance should be given two separate annotations, marking it as both an accomplishment and an achievement. I submit that these two descriptions are formally equivalent.

This link provides a neat solution to the relative complexity of having multiple annotations per clause in the corpus, the broadly unconstrained annotations made possible by the permissive interface of the annotation tool, as well as irregular or changing annotation styles over the course of the annotation project. Each clause in the corpus can be associated with a single label that combines one or more classes taken from the hierarchy of figure 7.3; this is termed its *ambiguity class*.

Table 7.5 shows the number of verb instances annotated with each unique ambiguity class in the corpus. As the table makes clear, ambiguity classes work very well to describe the results of the annotation effort. Most classes contain only a single category, and only two verb instances cannot be assigned an ambiguity class with only one or two sub-labels. Five verb instances are marked `is???`, which indicates that the verb is valid and not stative, but that nothing else can be said about the verb's aspectual class; these are often cases of bare *machen* 'do' in a small clause, which do not convey enough information for a more specific determination to be made.

Although we end up with more ambiguity classes than the simplex aspectual classes that we started with, there are not too many new additions, due to the long-tailed distribution of the ambiguity classes. As we will see in section 8.1, this proliferation of labels can be solved to a large degree by filtering on a minimum frequency threshold.

As can be seen in table 7.5, the most common type of aspectual ambiguity captured in the corpus concerns verb instances underspecified for the extended/punctual distinction. The second most common type captures verbs that are ambiguous in context between an extended telic reading and an unbounded atelic reading. This class contains examples of degree predicates such as *austrocknen* 'to dry out' (cf. example 7.5); other verbs in this class (e. g. *wirbeln* 'whirl, spin') may be incremental verbs (i. e., those taking an incremental theme argument cf. section 2.4.7, which are classed as telic predicates by Krifka, 1998). Filip (2012) points out that some examples of these (e. g., 'eat') can nevertheless manifest as telic or atelic ('They ate in two hours/for two hours'), leading her to conclude that 'incremental verbs ... are unspecified for telicity' (p 1208).

I make the assumption that part B (a random sample) is the best representation of actual language usage available from this annota-

Verb	Label	Corpus count
betreuen 'look after'	unbounded	50
zusammenfinden 'come together'	extended change	50
bedeuten 'mean'	stative	50
aufblitzen 'flash'	punctual no change	16
gelten 'apply to'	stative	12
stehen 'stand'	stative	11
blinzeln 'blink'	punctual no change	9
beginnen 'begin'	punctual change	7
glauben 'believe'	stative	7
hinweisen 'indicate'	extended no change	6

Table 7.6: Most common aspectually unambiguous verbs in the annotated corpus.

tion study: Although it over-represents infrequent verbs, because of the stratification of the sample on verb frequency classes, it is a broader sample than part A, and should capture a reasonably accurate picture of the distribution of aspectual classes in modern German. By comparing the ranking given in table 7.5 with the distribution of annotated labels in part B, shown in figure 7.7, we can conclude that the complete corpus with all parts together is biased, with accomplishment, achievement, and semelfactive verbs over-represented.

7.3.8 *Polysemy and aspectual ambiguity*

Of the total 578 verb types in the annotated corpus, only 123 verb types have instances annotated with different aspectual classes. Thus, 78.7% of the verbs are aspectually unambiguous on the type level. Table 7.6 shows the most common of these. Note that the corpus attests several verbs that are only annotated as stative, such as *bedeuten* 'mean', *gelten* 'be valid', *stehen* 'stand', and *glauben* 'believe'.

The aspectually unambiguous verb types are to a large extent verbs that are very infrequent in the corpus; their mean corpus count is 1.66, with a standard deviation of 4.15. This is very low compared to the verbs that are aspectually ambiguous, which have a mean corpus count of 25.55 (standard deviation 23.95). A one-sided Welch's t-test (used to control for unequal variances) confirms this, revealing that unambiguous verb types do indeed have significantly fewer instances in the annotated corpus than ambiguous verb types have ($p \ll 0.001$). These low counts suggest that aspectually unambiguous verbs may be less common than what we see here suggests; it is likely that annotating further instances of some of these 'unambiguous' verbs would reveal that they do in fact belong to more than one aspectual class.

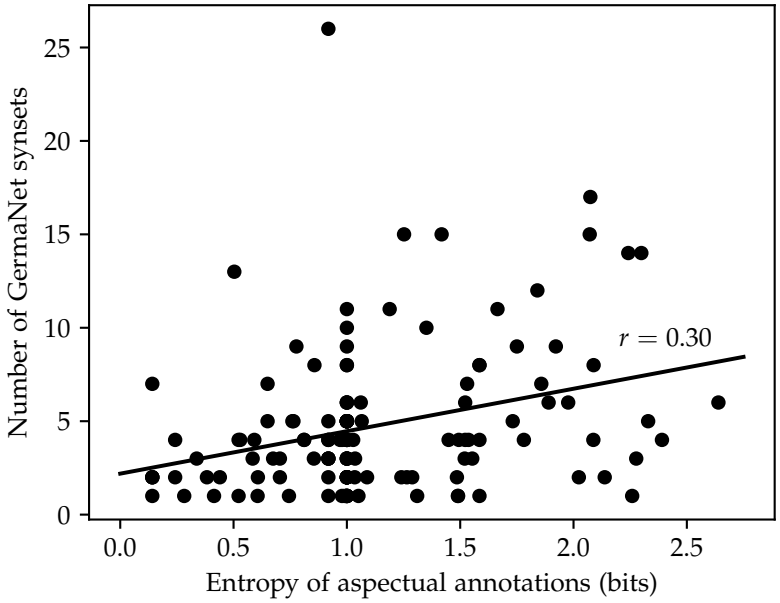


Figure 7.8: Aspectual ambiguity versus polysemy for verbs in the annotated corpus.

Note that the table shows that only three verbs from Part A are aspectually unambiguous. Conspicuously missing from this list is the only monosemous verb in Part A, *entstehen* 'arise'; while it is annotated as an accomplishment in 48 of 50 cases, it is also marked once as an achievement, and once as ambiguous between an accomplishment and achievement. Thus, this verb can be seen to have fallen victim to the problematic punctual/extended distinction.

Leaving aside the aspectually unambiguous verbs, let us take a look at the remaining 123 verbs. I can approximate the aspectual ambiguity of the verb as the entropy of the set of ambiguity class labels the verb has received in the corpus. I can also represent polysemy as the number of GermaNet synsets of a given verb. Figure 7.8 shows a regression analysis of these two variables. The trend is clearly visible: More polysemous verbs tend to be more aspectually ambiguous; this effect is statistically significant at $p < 0.001$. Similarly, an unpaired t-test corroborates this finding, indicating that unambiguous verb types have significantly fewer GermaNet synsets than ambiguous verb types ($p < 0.001$). This is in line with the observation made by Falk and Martin (2016) that verbs quite often have different readings that belong to different aspectual classes.

7.4 CONCLUSION

This chapter has described efforts to collect data on lexical aspect in German. At first, I set out to take the simplest and most direct path in this direction by translating the linguistic indicators of Siegel and McKeown to German and applying them to the SdeWaC corpus. However, on closer examination, this appears to be a dead end: The indicators occur only infrequently in my data, and, despite my best efforts to curate and filter, they do not seem to have a very high signal-to-noise ratio.

Upon this discovery, I switched tactics and mounted an effort to construct a manually-annotated corpus of German verbs labelled for features of lexical aspect; the corpus makes use of a novel typology of aspectual types, which is compatible with several pre-existing inventories. This annotated corpus, though small, has been constructed with considerable care; the quality of the corpus is borne out by the inter-annotator agreement measured in section 7.3.5. The annotation effort has underscored the prevalence of aspectual ambiguity in regular language; this accords both with predictions from semantic theory, as well as the empirical findings of Croft, Peskova and Regan (2016). I have also briefly investigated the interaction of verb polysemy and aspectual ambiguity, demonstrating that these two qualities are correlated. As we shall see in the next chapter, the distribution of aspectual classes found in the corpus matches up neatly with those reported in previous research in other languages; furthermore, the corpus is sufficiently accurate to be used as a knowledge source for applied NLP tasks.

The annotated corpus was released as a new freely available linguistic resource to the research community in Egg, Prepens and Roberts (2019); this publication also included preliminary results from some of the classifiers trained on the corpus in the next chapter. This corpus is the first computer-readable linguistic resource in German on lexical aspect.

8

COMPUTATIONAL ASPECT

With the successful completion of the aspectual corpus, I am now finally equipped with data on the aspectual behaviour of verb instances in the wild. In this chapter, I will describe work to produce a series of classifiers that can automatically label new unseen verb instances for aspectual class, in order to be able to study the aspectual behaviour of verbs in texts other than the annotated corpus, such as the remaining unannotated clauses in SdeWaC. The annotated corpus can be viewed as embodying a collection of aspectual knowledge about German verbs; to the extent that it is possible to automate the task of aspectually labelling verbs with a machine learning algorithm, the skill of the human annotators can be extracted from the corpus and transferred to other problems and domains. With the aid of the aspectual classifiers, I will then explore the impact of information about the aspectual structure of verbs on two applied NLP tasks.

The construction and training of the supervised classifiers are reported in section 8.1, along with an intrinsic evaluation. Section 8.2 presents an experiment to evaluate the aspectual classifiers extrinsically inside a semantic role labelling (SRL) task. Section 8.3 integrates the output of the aspectual classifiers into the verb clustering task that I have used in chapters 5 and 6. Finally, section 8.4 discusses the results of the experiments performed in the chapter and offers concluding thoughts.

8.1 AUTOMATIC ASPECTUAL CLASSIFICATION

The classifiers introduced in this section will use the annotated corpus developed in chapter 7 as a data set for training and testing with 10-fold cross validation. In *k-fold cross validation*, the labelled data points are partitioned into k different equally-sized groupings. Following this, the experiment is repeated k times; on each such *fold*, a different grouping of data points is used as the test set, and the $k-1$ remaining groupings are lumped together to form the training set. After all k iterations, the classifier will have been tested on the entire labelled data set, without ever having been tested on data that it had previously seen during training. The folds used here are generated so that they are *stratified* on the labels of the data points; this ensures that each of the k groupings contains roughly the same distribution of classes.

I make use of the property of the aspectual class hierarchy that it can be decomposed as discussed in section 7.3.1. This allows the prob-

lem of automatically labelling verb instances for their aspectual class to be broken down into a series of simpler classification tasks, both to be able to look at learning different aspectual features in isolation, as well as to allow comparisons to be made with previous research. The particular classification tasks are described later on in the coming subsections.

In all cases the classifiers in this section are built using a multi-class maximum entropy classifier with L2 regularisation ($\lambda^{-1} = 2.78$), like I have done previously with the edge labeller in section 5.6.1.[1] Evaluation is performed using the standard IR method (cf. section 3.7), comparing the predicted labels on the test set items to the ground truth labels produced by the human annotators. I aim here in several cases to establish direct comparisons with prior work, which is described in detail in section 4.4; for this reason, I will use accuracy as the statistic to measure the quality of a classifier.

The aspectual classifiers developed in this section take as features morphological, syntactic, and lexical properties of the verb and its clause, derivable from the automatic parse of the sentence, including:

- the POS of verb;
- the tense of the verb (e. g., past perfective, simple past, present, future);
- whether the verb is in the passive voice;
- whether the verb is finite;
- the verb's subject and object, and whether these are in the plural;
- the adverb closest to the verb in the clause, if present;
- the type of the clause the verb is embedded in (i. e., one of `i`, `S-2`, `S-dass`, `S-ob`, or `S-w`); and,
- the subcategorisation frame of the verb, as reported by the SCF tagger of section 5.2.

Classifiers are also given features indicating the presence of aspectual indicators on the verb as described in section 7.2. Finally, the classifiers were trained using external features drawn from other NLP resources. One of these features was the word vector of the verb, drawn from some word embedding model; I will go into greater detail in section 8.1.8. Another such external feature was a categorical label containing the semantic class listed in GermaNet for the verb, its subject (semantic head word), and object head. Semantic classes in GermaNet indicate from which semantic field a lemma is drawn from;

1 On these aspectual labelling tasks, I also tried random forest classifiers (Breiman, 2001) and XGBBoost (Chen and Guestrin, 2016), without great success.

Task	Baseline	Classifier	Classes	RER (%)
`full`	29.5	71.2	10	59.1
`egg`	44.5	78.5	7	61.3
`vendler`	36.8	73.0	7	57.3
`stative`	71.9	87.7	3	56.2
`telicity`	44.3	81.7	3	67.1
`culminated`	61.8	85.6	3	62.3
`extended`	80.8	88.1	3	38.0
`change`	53.2	85.2	3	68.4

Table 8.1: Classifier accuracies in percent on aspect labelling tasks.

examples for nouns include `Mensch` 'human', `Tier` 'animal', `Substanz` 'substance', `Ort` 'place', `Geschehen` 'event', etc.

I also experimented with adding 'Siegel & McKeown vectors'. This means, for each verb type, I build a vector of counts over the number of times that verb is observed in SdeWaC with each of the aspectual indicators introduced in section 7.2; however, this led to drops in performance (1–9% absolute accuracy) under every condition I tried.

On each task, the inherent difficulty of the classification task is tracked by evaluating the output of a simple baseline method. In this section, this baseline always predicts the most frequent label in the training set, ignoring any features provided during training. 'Classes' shows the number of labels that the classifier is trained to distinguish; a 7-way classification task is typically more difficult than a 3-way task. The column labelled 'RER' shows the Relative Error Reduction, the complement of the ratio of the number of data points mis-labelled by the trained classifier compared to the number mis-labelled by the baseline method, given in percent.

The classification tasks are described below. Results of the classification tasks are summarised in table 8.1. All accuracy scores achieved by the classifiers are significantly better than the corresponding baseline values as judged by McNemar's test ($p \ll 0.001$). Discussion follows in section 8.1.10.

8.1.1 *Full classifier*

The first task investigated here, called `full`, attempts to label verb instances according to the full taxonomy of aspectual classes depicted in figure 7.3. This is the most complicated classification task possible with this dataset, as it makes the most fine-grained distinctions and targets the greatest number of class labels.

The annotated verb instances from the aspectual corpus are transformed into data points labelled with ambiguity class labels, following the idea developed in section 7.3.7. As mentioned, the distribution of ambiguity classes (shown in table 7.5) is long-tailed, so that many of the uncommon and overly-specific data point labels can be removed with a simple filter on frequency. This makes the classification task significantly easier and enables better performance from the classifier. In this and the following two tasks, I define the problem of classifying to ambiguity classes by dropping data points that have ambiguity class labels seen fewer than 10 times in the aspectual corpus. For the `full` classification task, this filtering causes 40 data points to be dropped; the remaining verb instances are associated with a total of 10 different ambiguity class labels.

Thus, the `full` classifier tags verb instances with one of the six classes of the aspectual classification presented in section 7.3.1, plus four common ambiguity classes. The classifier achieves an accuracy of 71.2% on this 10-way classification task, over a baseline accuracy of 29.5%.

8.1.2 *Egg classifier*

The second task, `egg`, is a simpler labelling problem, because it drops the punctual–extended dichotomy, and hence makes fewer aspectual distinctions than the full classifier. The aspectual class taxonomy, when the punctual–extended category is dropped, corresponds to the taxonomy of Egg (2005).

As before, the resulting ambiguity classes are filtered with a frequency threshold of 10, causing 18 data points from the annotated corpus to be dropped, and leaving seven different ambiguity classes for the trained model to distinguish.

The `egg` classifier tags verb instances with one of Egg (1995)'s aspectual classes (i. e., stative, unbounded, bounded change-of-state, or bounded no-change-of-state), plus three common ambiguity classes. The classifier achieves an accuracy of 78.5%, over a baseline of 44.5%.

8.1.3 *Vendlerian classifier*

The third classification task drops the change–no-change distinction, which corresponds to the aspectual taxonomy of Vendler (1957). Filtering causes 26 data points to be dropped, and produces a data set containing seven distinct ambiguity classes.

This experiment can be roughly compared to the study done by Zarcone and Lenci (2008), who trained a classifier on a manually annotated dataset of Italian verb instances labelled for their Vendlerian classes (four classes). In comparison, my `vendler` classifier, which tags verb instances with one of Vendler's four aspectual classes (i. e.,

stative, unbounded, punctual, or extended), plus three common ambiguity classes, achieves an accuracy of 73.0% over a baseline of 36.8%. Both the system output and the baseline number are much lower than Zarcone and Lenci's reported accuracy of 85.4% over a baseline of 79.8%. Perhaps the reason is that they do a four-way classification, while here I perform a seven-way classification to Vendlerian ambiguity classes. Certainly, their baseline method is cleverer than mine, as it reports the most common label for each verb, instead of the most common class across the training data. However, it seems that there is some significant kind of difference in the setup: Where Zarcone and Lenci train a classifier to detect stativity, they achieve 92% accuracy over a baseline of 88% (high compared to my 88% over 72%); and a classifier they build to detect telicity attains 90% accuracy over the baseline 84% (where I calculate 82% over 44%).

Another difference is visible in the proportions of the classes. On this task, my dataset contains 37% accomplishments, 27% statives, 17% achievements, 15% activities, 2% verb instances ambiguous between accomplishments and achievements, 1% accomplishment/activities, and 1% accomplishment/statives. This distribution differs in several ways to Zarcone and Lenci's reported breakdown of 41% achievements, 26% accomplishments, 18% statives, and 13% activities, constructed as a verb sample on 28 verb types. Notably, my corpus contains many more accomplishments and many fewer achievements, and I also count slightly more statives than Zarcone and Lenci do.

On these first three tasks, the relative error reduction over the baseline is similar at ~60%. The `egg` classifier shows better performance than the `vendler` classifier, and the baselines for these two tasks show the same effect. This finding could be offered as evidence that the change–no-change distinction is more well-founded than the punctual–durative distinction; certainly, the conclusion based on this evidence must be that judging whether a verb involves a change of state is an easier task than judging whether it is extended.

8.1.4 *Stativity task*

The next three classification tasks reduce the labelled corpus to simple distinctions, for better comparison with previous research.

The `stative` task defined in this section follows Friedrich and Palmer (2014), who produced a manually annotated dataset of English verb instances labelled as stative (17%), dynamic (73%), or ambiguous in context (10%). They then trained and tested classifiers on this dataset.

To recreate this experimental design with my annotated corpus, I can discard aspectual annotations for all categories except stativity. Transforming the annotated corpus in this way gives 1,077 verb instances that are stative (27%), with a further 2,915 that are dynamic (72%); 60 are ambiguous in context (2%). The class proportions agree

fairly well with those reported by Friedrich and Palmer, although they have fewer statives and more ambiguous verbs than are found in my aspectually annotated corpus.

Using the ten-fold cross-validation setup of Friedrich and Palmer's Experiment One, my `stative` classifier, which tags verb instances as `stative`, `dynamic`, or `ambiguous` in context, achieves 87.7% accuracy over the baseline 71.9%. The baseline is in the same range as Friedrich and Palmer's reported 72.5%, and my classifier attains a higher accuracy than their random forest's 84.1% accuracy.

Cross-validating while stratifying the folds on verb lemma, which duplicates the design of Friedrich and Palmer's Experiment Two, my classifier achieves 81.1% accuracy, about on par with Friedrich and Palmer's 81.9%.

The first experiment of Siegel and McKeown (2000) is also of identical design to Friedrich and Palmer's Experiment One (indeed, this isomorphism was intentional); however, both Siegel and McKeown's baseline accuracy of 83.8%, as well as their decision tree's evaluation score of 93.9% are much higher than the scores reported by Friedrich and Palmer or than I can report here. Note that Siegel and McKeown conducted their first experiment on a labelled corpus drawn from a highly specific textual domain, namely medical reports.

8.1.5 *Telicity task*

The fifth task replicates the study done by Falk and Martin (2016), who trained classifiers on a dataset of French verbs labelled as atelic (35%), telic (48%), or of variable telicity (16%).

To map my annotated corpus to the scheme used by Falk and Martin, I class stative and unbounded verbs together as atelic (1707 instances, 42% of the corpus), and change-of-state verbs as telic (1794, 44%); no-change verbs, as well as any instances ambiguous in context between atelic and telic are classed as having variable telicity (551, 14%). The proportions of the classes agree well with the distribution of the high level groupings described by Falk and Martin.

My `telicity` classifier, which tags verb instances as `telic`, `atelic`, or `variable` in context, attains a 81.7% accuracy over a baseline of 44.3%. The baseline is comparable to Falk and Martin's value of 48.4%, and my accuracy is better than their 67.5%.

8.1.6 *Culmination task*

The sixth classification task recreates the second experiment of Siegel and McKeown (2000), which trained classifiers on a dataset of English verbs annotated as non-culminated (37%) or culminated (63%).

To duplicate this setup, I take culminated verb phrases to be VPs with change-of-state verbs, and non-culminated VPs have either a no-

change verb or an unbounded verb. I include a third category to catch those verb instances that are ambiguous in context for culmination. This mapping gives 1,834 culminated verbs (62%) and 1,077 non-culminated verbs (36%); 59 are ambiguous (2%). These proportions neatly match those reported by Siegel and McKeown.

My baseline method attains 61.8% accuracy, similar to Siegel and McKeown's reported 63.3%, and my `culminated` classifier, which tags verb instances as `culminated`, `non-culminated`, or `ambiguous` in context, achieves better performance with 85.6% accuracy than their reported 74.0%.

8.1.7 *Extended and change tasks*

Finally, the `extended` and `change` classifiers mirror the `stative` classifier by focusing on a single category and ignoring all others. I developed these for completeness; there are no previously published examples of such aspectual classifiers to give points of comparison.

For the `extended` classifier, the verb instances from the corpus are transformed to give a dataset containing 3,272 extended verb instances, 693 punctual instances, and 87 that are ambiguous for the punctual-extended distinction in context. The classifier, which tags verb instances as `extended`, `punctual`, or `ambiguous` in context, achieves an accuracy of 88.1% over a baseline of 80.8%. The high baseline value observed here is indicative of the highly skewed instance labels, almost all of which are extended. This classifier achieves the lowest Relative Error Reduction of all the conditions tested; this may simply be a result of the unbalanced dataset, or it may indicate that it is more difficult for the maximum entropy learner to come to terms with the punctual-extended distinction than the other categories.

For the `change` classifier, verb instances are again transformed, giving 2,154 no-change events, 1,794 change events, and 104 ambiguous events. The `change` classifier tags verb instances as `change`, `no-change`, or `ambiguous` in context, and achieves an accuracy of 85.2% over a baseline of 53.2%. This posts the highest Relative Error Reduction of all the conditions tested here, indicating that this category is somehow amenable to statistical modelling of the kind I have performed here.

8.1.8 *Word embeddings*

As introduced in section 3.2.1, many modern NLP applications make use of word embedding vectors to crystallise some sort of semantic representation out of raw text from a domain of interest. The vector representation, a dense collection of a few hundred numerical values, can be easily passed to a machine learning classifier, and using word vectors as features should help a classifier to better deal with unseen

verb instances. To this end, I produced several word embedding models and also collected two pre-built models, and tried various ways of integrating these into the classifiers described above. These embeddings are:

`wikipedia` Trained on the text of the German Wikipedia accessed in April 2015, cleaned of MediaWiki markup and tokenized using a few regular expressions in Perl and Python, unlemmatised, using `word2vec` to produce 400-dimensional CBOW word vectors, window size 5, 10 negatives, sampling 10^{-5};

`wikiextract` as with `wikipedia`, but cleaned of markup using WikiExtractor[2]; unlemmatised, `word2vec` settings as before;

`wikipedia-lemmas` as with `wikipedia`, but lemmatised by the TreeTagger (section 3.1.2); `word2vec` settings as before;

`sdewac-lemmas` word vectors built on the SdeWaC corpus, lemmatised with the `mate-tools` parser, and using `word2vec` with the settings recommended by Baroni, Dinu and Kruszewski (2014): 500 dimensional CBOW vectors, with a window size of 5 and 5 negatives, sampling 10^{-3};

`spacy` 384-dimensional word vectors trained on the TIGER corpus and Wikipedia; provided by the SpaCy project[3];

`fasttext` 300-dimensional CBOW word vectors with character *n*-grams of length 5, a window of size 5 and 10 negatives, trained on Common Crawl and Wikipedia; provided by the FastText project[4] (Grave et al., 2018).

The best performance from the aspectual classifiers was obtained using my `sdewac-lemmas` model, and all evaluation numbers reported in this chapter make use of this model. The `fasttext` vectors were almost as good on all the tasks I tried; they have the advantage of coming off the shelf and pre-trained. `wikipedia-lemmas` was less helpful than `sdewac-lemmas`, and sometimes no better than not having an embedding vector at all. `wikipedia` and `wikiextract` were even less helpful, and suffered from missing lexical items because they were built on unlemmatised text. The `spacy` vectors reduced performance.

8.1.9 *Feature importances*

I investigated which features were most highly weighted by the trained maximum entropy classifiers. All classifiers attend quite strongly to bag-of-words features and particular dimensions in the `sdewac-lemmas`

2 http://attardi.github.io/wikiextractor/
3 https://spacy.io
4 https://fasttext.cc/

embedding vector for the verb, and most pay careful attention to the preposition included in the SCF, with special interest shown to PPs with accusative arguments (these often code for motion or transfer). Another pair of features popular with many of the classifiers are the adverbs *mehr* 'more' and *ernst* 'earnestly'. Many classifiers weight the subject word highly: *Jeder* 'everyone/anyone' is a favourite (perhaps indicating genericity), as are the simple pronouns *ich* 'I', *es* 'it' and *er* 'he' (possibly, these are more often agentive than not). The least important features appear to be voice (active/passive), tense, finiteness, and the POS of the verb.

The `change` classifier pays attention to the GermaNet semantic class of the subject (`natPhaenomen` 'natural phenomenon', `Mensch` 'kinds of humans', `Menge` 'quantities, collections', `Pflanze` 'plant'), verb (`Schoepfung` 'change events'), and object (`Gefuehl` 'emotion and perception'). `telicity` likes the GermaNet subject classes `natPhaenomen`, `Pflanze`, `Ort` 'place', and `Artefakt` 'artefact'. `culminated` is fond of the subject classes `natPhaenomen`, `Koerper` 'body parts, diseases', and `Pflanze`, as well as the adverbs *stark* 'strongly', *gut* 'well', *so* 'thus', and *schon* 'already'. `extended` looks at whether the direct object is plural, whether the SCF is intransitive, and likes the GermaNet subject classes `Substanz` 'substances' and `Nahrung` 'foodstuffs', as well as the adverbs *bereits* 'already', and *genau* 'exactly'. Finally, `stative` attends to the GermaNet subject classes `Mensch` and `Ort` and the object class `Gruppe` 'organisations', as well as the adverbs *mutig* 'courageously', *schlicht* 'unpretentiously', and *sogar* 'even'; it has also learned to disambiguate the verb *stapeln* 'pile, stack' on the basis of its SCF: `npr` is stative (to stand in a pile), whereas `na` is dynamic (to stack something up).

This qualitative examination has demonstrated the utility of sentential adjuncts (such as adverbial modifiers to the verb). The feature importances extracted from the `extended` classifier also seem to suggest that at least the plural marking of the verb's direct object is a strong predictor of the VP's aspectual class. It is likely that plural subjects can also have effects, but this should be less common, only being critical in cases of unaccusative verbs.

The failure of the Siegel and McKeown vectors to improve classifier performance accords with the findings of the limited in-vitro testing in section 7.2, and offers additional evidence that the straightforward translation of these indicators from English into German was unsuccessful.

That the SCF is a highly weighted feature for the classifiers (particularly `stative` and `extended`) demonstrates once again that subcategorisation information is useful for modelling verb semantics.

8.1.10 *Discussion*

This section has presented a number of experiments with supervised classifiers, with the evaluation indicating good performance. This enterprise has been materially assisted by the flexibility of the six-way aspectual classification introduced in section 7.3.1, which permits a single annotated resource to be used in different ways and for different tasks. I have tried here to compare to, or replicate the experimental setups of various prior research projects. Of course, the annotated datasets created and used in these papers are in English, Italian, and French, and cannot be thought of as 'the same thing' as the German data I use here; in some cases, e. g., Siegel and McKeown's stativity experiment, the corpora are constructed for highly specific domains. Nevertheless, there are many cases where baseline accuracy, inter-annotator agreement, or labelled class proportions line up very neatly against my results; there are also several cases where my classifiers show results consistent with or better than those previously reported.

These observations encourage a positive view of the quality of the aspectual corpus: The good performance of the classifiers seems like it must be a result of the careful annotations. Furthermore, the relatively high accuracies achieved by the classifiers on some of the tasks reported in this section make it seem like these classifiers might be used for real-world tasks, even though the aspectual corpus only contains some 4,100 labelled data points. This would enable a kind of transfer of information from the annotated corpus to new applications and domains. The following sections will attempt to do exactly this.

8.2 ASPECT FOR SEMANTIC ROLE LABELLING

This section presents an extrinsic evaluation of the aspectual classifiers developed up until now; by extension, it can also be seen as a convoluted extrinsic evaluation of the annotated corpus. The specific applied problem I address here is semantic role labelling.

Semantic role labelling is a common NLP task, and one which is intimately connected to the semantics of the verb and its arguments. The goal of SRL is to automatically determine the thematic role for each syntactic concomitant of a verb instance, thus deciphering the verb's argument linking. While I am not aware of any previous research looking at the intersection of SRL and the lexical aspect of verbs, there are reasons to think that aspect features of the verb could be relevant for the task. Apart from overt connections between aspect and thematic roles, such as Incremental Theme[5] and other effects of the

5 Recall the emphasis placed by Krifka and others on the category of Incremental Theme for determining whether a phrase is telic or not, cf. section 2.4.7.

Patient on aspectual class, a general understanding of the intended sense of the verb in context should be informative when trying to decide how the verb's arguments are involved in the action.

For this experiment, I use the `mate-tools` SRL labelling system of Björkelund, Hafdell and Nugues (2009)[6]; this was the second-best system overall in the CoNLL 2009 ST closed challenge semantic labelling task (Hajič et al., 2009), and the best system for German, with a published semantic F_1 score of 79.71. An extension of `mate-tools` called `mateplus` (Roth and Lapata, 2015)[7] today holds the current record for best semantic F_1 score in German on this test set, at 81.38.

The `mate-tools` software uses a pipeline of maximum entropy classifiers to perform, in order:

1. predicate identification, `pi`;
2. predicate disambiguation, `pd`;
3. argument identification, `ai`; and
4. argument classification, `ac`.

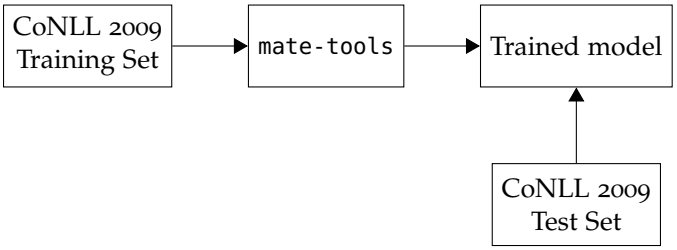
This is followed by a reranking stage, which improves performance by estimating the likelihood of a global combination of classifier outputs, using a beam search. This can allow, for example, a particular pipeline stage to produce a less-likely labelling that, in turn, enables a more-likely labelling of a later pipeline stage.

I use the CoNLL 2009 Shared Task (cf. section 3.4) for evaluation, training the SRL system on the training data using the default parameters, and testing on the test data in SRL-only mode (i. e., the SRL labels are predicted based on the gold standard parses, not on automatically-predicted syntactic analysis). Under the experimental condition, I provide additional features to the SRL system during training and testing; these features capture information about the predicate's aspectual class. Because predicate identification is not needed in SRL-only mode, the experimental condition does not modify the `pi` stage of the SRL pipeline. The experimental procedure is diagrammed in figure 8.1.

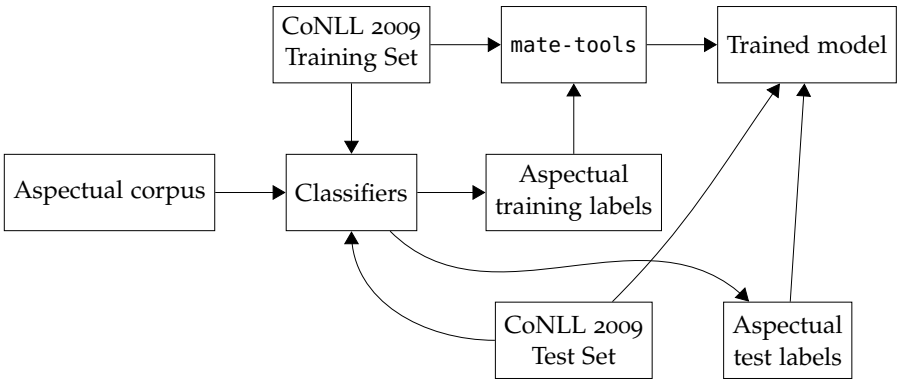
The additional features used in the experimental condition are the labels predicted by the classifiers described in section 8.1 and summarised in table 8.1, with the modification that the classifiers are trained on the full aspectually labelled corpus, and no labelled data are held out for testing. The best-performing feature set is used, including the `sdewac-lemmas` embedding vectors.

These classifiers were used to predict labels for all predicates in the CoNLL 2009 Shared Task training and test data. Each classifier reports both the most likely label, as well as the estimated probability

6 https://code.google.com/archive/p/mate-tools/
7 https://github.com/microth/mateplus



(a) Control condition.



(b) Experimental condition.

Figure 8.1: Setup of SRL experiment.

Quantile	Confidence value	Semantic F_1
0	—	80.21
0.01	0.454	79.91
0.025	0.501	79.92
0.05	0.538	80.20
0.1	0.623	79.96
0.2	0.785	79.70
0.3	0.910	80.01
0.4	0.962	79.90
0.5	0.981	79.79

Table 8.2: Semantic F_1 of the SRL system (experimental condition) as a function of a threshold in confidence value for the automatically-produced aspectual labels. Quantiles are given as decimal expansions of fractions, so that removing labels with $p < 0.981$ would replace half of all the aspectual label data with the `unknown` label.

	Semantic F_1
Control	79.31
With aspectual features	80.21

Table 8.3: Evaluation of the effect of adding aspectual label data to the `mate-tools` SRL system on the CoNLL 2009 shared task.

of the label, which can be seen as a proxy to the classifier's confidence in its prediction.

Under the experimental condition, the SRL system used all eight automatically predicted labels as categorical features in its internal pipeline and reranker. In development, no particular combination of the eight labels outperformed the use of all eight together; further, no important interactions were observed between the aspectual label features, and the other (lexical, morphological, and syntactic) features used internally by the SRL system.

Exploration of the predicted probabilities of the aspectual labels proved fruitless, as shown in table 8.2. Replacing aspectual labels with a generic `unknown` label when their predicted probability is below a certain threshold appeared to reduce performance overall.

As shown in table 8.3, the experimental condition using automatically predicted aspectual features improves performance of the `mate-tools` SRL system on the CoNLL 2009 shared task by $+0.90\%$ F_1 score absolute, or a Relative Error Reduction of 4.3%. The statistical significance of this difference was estimated using a randomisation test (Yeh, 2000). In this setup, the null hypothesis asserts that both the

control method and the experimental method are equally good at labelling semantic roles. Under this assumption, their output should be more or less equivalent. Thus, the test randomly shuffles the outputs for particular sentences between the control and experimental conditions. The two permuted outputs are scored using the official CoNLL 2009 scorer, and the difference in scores taken; this shuffle-and-score routine is then repeated many times. Statistical significance can be estimated by comparing the count nc of the number of times this simulated score difference meets or exceeds the actual observed score difference of 0.90, to the total number of permutations sampled nt:

$$p < \frac{nc+1}{nt+1}$$

I can report $nc = 165, nt = 10000, p < 0.0166$, which is significant at the $p < 0.05$ level.

8.3 ASPECT FOR VERB CLUSTERING

In chapters 5 and 6 we have seen two experiments using automatic verb clustering to evaluate descriptions of verb argument structure. In this section, I repeat this experiment and extend it to include automatically predicted aspectual features of verb instances. The working hypothesis here is that verbs belonging to the same verb class will tend to share not only argument structure but also aspectual structure.

As before, I represent verbs by vectors representing their subcategorisation preferences; the verbs can then be partitioned into classes using hierarchical clustering with Ward's criterion and the Jensen-Shannon divergence as the distance measure. The PairF evaluation metric is used to compare the automatically-generated clustering to the gold standard of Schulte im Walde (2006). This time, I will parameterise the verb's SCFs with selectional preferences as developed in chapter 6 and/or the output of one of the aspectual classifiers developed in section 8.1.

This experiment is performed by counting verb-SCF co-occurrences on the full SdeWaC corpus: For the 168 verb types used in the gold standard, this gives 10,118,405 verb instances. After all co-occurrences have been collected, counts lower than a threshold of two are removed to regularise the data and reduce the computer memory required for clustering.

Table 8.4 shows the results; the highest PairF value in each row is printed in bold face. The none baseline column represents the condition where the SCF is not parameterised for any aspectual features. The other columns indicate conditions making use of the classifiers from section 8.1; as in section 8.2, these are trained on the full trained on the entire aspectually labelled corpus, with no data held out.

I compare several models of SCF and selectional preferences:

	none	stative	change	culm.	tel.	ext.	egg	vendler	full
LP5K	**36.02**	33.15	29.67	28.28	29.78	29.27	25.51	21.59	19.89
LP10K	**35.28**	33.52	30.48	30.99	29.41	31.27	25.97	22.65	20.03
LP1K	31.31	**35.13**	29.62	28.89	29.19	25.48	25.27	21.08	19.04
SUN10K	30.92	**34.81**	31.76	33.19	30.75	27.97	26.09	23.20	21.01
WSM10K	32.24	**33.52**	29.39	29.05	28.29	25.76	23.90	21.58	20.48
SCF	25.59	**27.00**	26.85	26.94	23.18	23.98	19.07	19.26	17.44
No SCF	3.94	6.33	6.67	6.90	7.68	5.18	8.63	8.51	**9.12**
Mean	27.90	**29.07**	26.35	26.32	25.47	24.13	22.06	19.70	18.14

Table 8.4: PairF values from verb clustering evaluation of SCFs parameterised for automatically labelled aspectual features.

- LP5K, LP10K, and LP1K indicate the lexical preferences model from section 6.2.1 with $N = \{5000, 10000, 1000\}$ nouns, respectively;
- SUN10K indicates the SUN model from section 6.2.2 with $N = 10000$ nouns in $M = 1000$ concept clusters; and
- WSM10K is the word space model from section 6.2.3 with $N = 10000$ nouns in $M = 500$ concept clusters.
- The SCF model represents a baseline condition with only SCFs and no selectional preference information.
- Finally, the 'No SCF' condition shows verb clustering performance using **only** the aspectually predicted features; that is, each verb is represented by its probability distribution over the aspectual categories predicted by the supervised classifier. The `noscf-none` condition contains exactly zero information (that is, every verb is represented by the same unit length single-dimensional vector); this is why the PairF value of 3.94 obtained here is so close to the random baseline value of 2.08.

The 'Mean' row displays the arithmetic mean along each column.

The best performing models overall are the LP5K and LP10K models; these are the only models that do not benefit from additional information about lexical aspect. All other models are improved by the automatically predicted aspectual features, although there is a clear pattern in that some of the classifiers are more helpful than others. The `stative` classifier improves verb classification performance on all models except LP5K and LP10K. The WSM10K, LP1K, SUN10K, and SCF models achieve their best performance with the addition of this feature. Under the 'No SCF' condition, we can see that verb clustering performance increases as the classifiers provide more detailed predictions. The best performance is obtained using the `full` classifier, which classifies verb instances into one of 10 ambiguity classes.

	Coef.	Std. err.	t	$P > \|t\|$	
SCF `none`	26.81	1.27	21.16	0.000	**
Model					
LP5K	4.87	1.22	3.98	0.000	**
LP10K	5.59	1.22	4.57	0.000	**
LP1K	3.97	1.22	3.24	0.002	**
SUN10K	5.60	1.22	4.57	0.000	**
WSM10K	3.88	1.22	3.17	0.003	**
No SCF	−16.26	1.22	−13.28	0.000	**
Classifier					
`stative`	1.17	1.39	0.84	0.405	
`change`	−1.55	1.39	−1.12	0.269	
`culminated`	−1.58	1.39	−1.14	0.261	
`telicity`	−2.43	1.39	−1.75	0.086	
`extended`	−3.77	1.39	−2.72	0.009	*
`egg`	−5.84	1.39	−4.21	0.000	**
`vendler`	−8.20	1.39	−5.91	0.000	**
`full`	−9.75	1.39	−7.03	0.000	**

$*: P < 0.01; **: P < 0.005$

No. Observations:	63	Prob(Omnibus):	0.001
Df Residuals:	48	Skew:	0.703
Df Model:	14	Kurtosis:	6.075
Adj. R-squared:	0.906	Durbin-Watson:	1.571
F-statistic:	43.70	Jarque-Bera (JB):	30.016
Prob (F-statistic):	1.86×10^{-22}	Prob(JB):	3.03×10^{-7}
Omnibus:	15.062	Cond. No.	10.5

Table 8.5: Ordinary least squares analysis of the verb clustering experiment.

To assess the statistical significance of these results, I performed an ordinary least squares linear regression, the results of which are shown in table 8.5. Under this analysis, the verb clustering score is modelled as the sum of an effect due to the choice of selectional preference model and an effect due to the choice of aspectual classifier (PairF $\sim$ Model + Classifier).

All selectional preference models perform significantly better than the baseline SCF-only condition, except the 'No SCF' condition, which achieves significantly worse scores. The `stative` classifier delivers better performance than the `none` baseline, although this difference is not statistically significant. By contrast, the `extended`, `egg`, `vendler`, and `full` classifiers deliver significantly worse performance than the baseline.

The adjusted R^2 value shows that the choice of Model and Classifier together explain 90.6% of the variance seen in the PairF scores, indicating that the model has very high goodness of fit, and the F statistic reveals that the linear model captures statistically significant effects (that is, we must reject as exceedingly unlikely the null hypothesis that the regression coefficients are only non-zero due to chance).

Linear regression models the dependent variable as a linear function of the independent variables; deviations in the dependent variable from the predicted linear function of the independent variables, called '*residuals*', are seen as error in the model, and are modelled as a random variable. The analysis places some assumptions on the behaviour of this random variable: Error residuals should be both *homoscedastic* (the residuals all have the same variance) and linearly independent (the residuals must not be correlated with each other); furthermore, error residuals should also be normally distributed.

The Durbin-Watson statistic is between 1 and 2, indicating that the variance of error residuals is constant, and the condition number is less than 15, indicating that the residuals are relatively independent. However, the residuals are not normally distributed. This is indicated by the low Omnibus and Jarque-Bera probabilities; more specifically, we can see that the distribution of residuals is moderately skewed (skew is larger than 0.5), and the kurtosis is larger than 3, indicating that the distribution of residuals is narrow with long tails.

These departures from the assumptions of the linear regression somewhat reduce the validity of this statistical model; while the model makes strong predictions about verb clustering performance, we should remain cautious in our interpretations. Transforming the PairF score with the natural logarithm improved the skewness of the residuals but further increased kurtosis, giving worse scores on the tests of normality and indicating that this modification to the model is no better than the linear version.

To investigate the effect of sparsity on verb clustering performance, figure 8.2 shows the PairF results from table 8.4 plotted against the

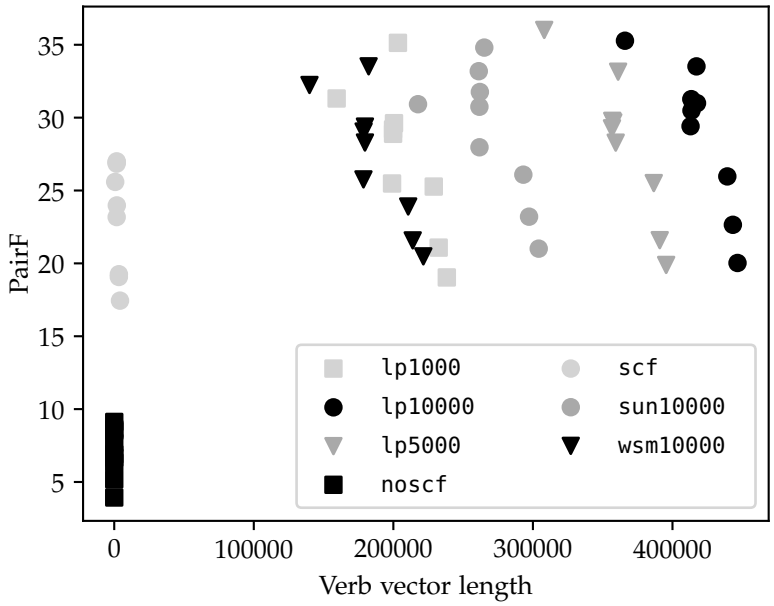


Figure 8.2: Scatterplot of verb clustering performance by vector length.

length of the parameterised verb vectors. On each condition, the vector length increases with the more detailed classifiers, so that the `none` condition (0 aspectual classes) will be placed to the left and the `full` condition (10 aspectual classes) is to the right. The downward trend visible inside each condition is due to the lower accuracy of the more detailed classifiers. Overall, the graph does not suggest that data sparsity is limiting verb clustering performance under these conditions. Rather, the more apt conclusion here is that the aspectual classifiers are only of limited use on this task.

From this point of view, the observation that the LP5K model does not improve when aspectual features are added to it could be taken to indicate that the model already captures so much high quality information about aspectual class through its modelling of the predicate-argument structure of verbs that the automatic classifiers of section 8.1 can only reduce its performance. If this were to be the case, we might expect that the LP5K model would be useful itself for automatic aspectual labelling.

To test this hypothesis, I repeat the ten-fold cross-validation experiment predicting the `stative` category from section 8.1.4, representing each verb type in the annotated aspectual corpus by its subcategorisation preferences vector as represented by the LP5K model; like I have done so far in this section, I drop parameterised SCF dimensions with fewer than two total co-occurrences in the SdeWaC data. Because this method uses only type-based features and no token-based features I will group the folds by verb type (i. e., Friedrich and Palmer 2014's

Classifier	Accuracy	Parameters
LP5K Maximum entropy	76.4	L1 regularisation, $\lambda^{-1} = 100$
LP5K SVM	76.0	linear kernel, $C = 1$
LP5K Random forest	75.3	100 estimators
`sdewac-lemmas` Maximum entropy	79.9	L2 regularisation, $\lambda^{-1} = 0.1$
`sdewac-lemmas` SVM	83.9	RBF kernel, $C = 1$
Baseline	71.9	Most likely class
F&P Expt. 2	81.9	
`stative`	81.1	

Table 8.6: The LP5K model used as classification features for the `stative` task.

Experiment Two), so that the verb types in each test fold are not attested in any of the corresponding training folds. This ensures that the classifier is never tested on a verb type that it has seen during training.

I trained a small number of supervised classifiers; the results of this are summarised in the top part of table 8.6. All the methods were about equally successful, with maximum entropy performing best. For comparison, the bottom part of the table shows the most likely class baseline, and the accuracies attained by Friedrich and Palmer (2014) and the `stative` classifier I developed in section 8.1.4. While the LP5K model performs significantly better than the baseline, it achieves significantly worse scores than the `stative` classifier. The middle part of the table shows another pair of classifiers using verb type information, this time using the 500-dimensional word embedding vectors I have been using in this chapter. The `sdewac-lemmas` model is significantly better than the baseline, and the SVM classifier trained on these embeddings also performs significantly better than the `stative` classifier.

This small experiment has failed to demonstrate that the LP5K model contains enough information about verb types to match the `stative` classifier's performance. Still, the very fact that the LP5K model outperforms the most likely class baseline demonstrates that it does encode information from which the aspectual category of stativity can be predicted. Likewise, this study has shown that high-quality word embedding vectors also capture properties of verb meaning that are useful for predicting aspectual features.

8.4 CONCLUSION

Section 8.1 has described the development of a series of automatic classifiers for lexical aspect. In measuring the performance of these

classifiers, I have been able to demonstrate reasonably good performance in intrinsic evaluations, despite the size of the aspectual corpus, which contains only slightly more than 4,000 labelled verb instances. In particular, I have been able to show that the classifiers perform to a large degree comparably with the state-of-the-art in other languages as reported in previously published research. The remainder of the chapter then makes use of these classifiers to augment other NLP applications.

In section 8.2, I demonstrated that an automatic SRL system could benefit from even automatically-predicted aspectual features. To my knowledge, this is the first use of computational aspect for SRL.

Section 8.3 repeated the automatic verb clustering experiment that I have used in previous chapters, demonstrating that automatically-predicted aspectual features can improve verb clustering on top of features capturing argument structure; the very best argument structure models, however, exhibited decreased performance with the added aspectual features. I believe this to be the first time computational aspect has been employed for automatic verb classification. A final experiment demonstrates that a model of argument structure is able to outperform a baseline method on an aspectual labelling task.

The results presented in this chapter suggest that argument structure and aspectual structure information are not orthogonal and that there is some degree of overlap between these facets in the empirical data derived from the SdeWaC corpus. The semantic role labelling experiment and the task of automatic verb classification based on the principle of diathesis alternation are both NLP tasks that clearly concern argument structure, and I have shown that automatically predicted aspectual features can deliver an improvement in performance on these tasks. Similarly, we have seen that an empirical model of argument structure can be used on an applied task that explicitly concerns aspectual structure, namely classifying verbs as being stative or dynamic. These results support the hypothesis that verbs from the same verb class will tend to have the same aspectual class.

The work in this chapter and the last provides circumstantial evidence for the hypothesis that the extended/punctual distinction is a problematic category of lexical aspect. In section 7.3.5 we saw that this category has low inter-annotator agreement, indicating that it is hard for humans to judge how long an event takes on the basis of a linguistic description. Additionally, it appears that durativity is the hardest aspectual feature to classify in the experiments that I have performed. For instance, the `egg` classifier delivers better performance than the `vendler` classifier in section 8.1; these two classifiers attempt tasks of similar difficulty, distinguishing seven classes each, but the `egg` classifier drops the extended/punctual distinction, whereas the `vendler` classifier drops the change/no-change distinction. Similar conclusions are supported by the automatic verb clustering experi-

ment of section 8.3: Here, the `extended` classifier is the least successful of the three-way classifiers I tried, and it is also the worst of all the classifiers under the 'No SCF' condition. By contrast, the competing change/no-change category performs much better in comparison, as seen by the good evaluation results attained by the `change` classifier. These findings agree with the observations of Zarcone and Lenci (2008), who found that their maximum entropy classifier had more trouble distinguishing punctual vs. durative verbs than stative vs. dynamic or telic vs. atelic.

Part IV

CONCLUSION

9 CONCLUSION

This dissertation has set out to empirically construct models of verb meaning based on information extracted from a large corpus of text.

Chapter 5 has explored the topic of subcategorisation acquisition for German. This work resulted in a large automatically-produced valency lexicon for German verbs, which was released to the research community.

Chapter 6 explored the selectional preferences of German verbs. I successfully developed an automatically-acquired model with very good performance, as measured on the verb clustering experiment. The best-performing model of selectional preferences induced noun classes (or concept clusters) that were relatively fine-grained, and were semantically specific, with pronounced synonym and co-hyponym structure. This accords well with predictions made by semantic theory.

Chapter 7 produced a manually annotated corpus of verbal aspect, which was released to the research community. This is the first machine-readable lexical resource on lexical aspect available for German. The corpus data reveal that verbs are frequently aspectually ambiguous (at least 20% of verb types can be expected to present with different aspectual classes in different contexts), and I am able to show a correlation between the polysemy of a verb type and its degree of aspectual ambiguity.

Chapter 8 developed a series of automatic classifiers for features of lexical aspect. These were then used to improve an automatic semantic role labelling system. They were also tested on the automatic verb classification task, with mixed results. To my knowledge, this is the first use of lexical aspect for semantic role labelling and automatic verb classification. Several empirical results presented in this chapter suggest that the facets of argument structure and aspectual structure provide overlapping descriptions of verb meaning. This connection could be due to the influence thematic roles have on aspect, where properties of the verb's Patient inform the aspectual type of the verb phrase; however, these results could also be taken as evidence that, in some more profound way, argument structure and aspectual structure describe the same fundamental phenomena.

The linguistic resources that I produced as a part of this doctoral project were described in publications that were presented at well-respected conferences of computational linguistics; hopefully, this will aid in disseminating them to a wide audience.

I released several bits of software code that I wrote in the course of this dissertation, including the `pygermanet` library for easily accessing information in the GermaNet database, and the TGrep2 phrase structure search engine, which has become a part of the NLTK project.

Of theoretical import are the findings that the punctual/extended distinction in lexical aspect is more difficult to annotate and less easy to model than other categories. This may be taken as a gentle recommendation against the unconsidered adoption of Vendler's aspectual classes for future work on lexical aspect.

Part V

APPENDIX

A

ASPECTUAL ANNOTATION GUIDELINES

These guidelines were written by Markus Egg, Helena Prepens, and Will Roberts.

A.1 INTRODUCTION

'Aspect' characterises the way in which the temporal progression of a state of affairs or *eventuality* is described (e.g., does it involve change, temporal boundaries, is it extended or punctual). Depending on differences in this description, verbs (sometimes only verbs together with specific arguments) can be grouped into a number of aspectual classes. It is the goal of the present annotation initiative to manually produce a database that offers such a classification.

You will be given a set of automatically segmented and parsed German sentences, whose main verb has been marked. It is your task now to annotate this verb token for its aspectual class in the context of the sentence.

In this initiative, we define the aspectual class of a verb by choosing one option each in four aspectual dichotomies, which are described in detail in appendix A.2.

- stative/dynamic
- bounded/unbounded
- change/no change
- punctual/durative

Sentences may also be classified as invalid, if they are unsuitable for the task of aspectual classification.

Our goal is annotating verb tokens in their respective context, not annotating the clauses or sentences they are part of. The reason for this is that clauses and sentences can additionally contain aspectually relevant phenomena (like durative or frequency adverbials) that can exert aspectual influence in that they can map the aspectual class of the verb onto a new aspectual class for the constituent comprising them and the verb.

We will explicate the exact procedure of getting this right below in appendix A.3. This includes an account of verbs for which a specific argument is relevant for aspectual classification, e.g., in the case of *eat an apple* (which introduces inherent boundaries) as opposed to *eat apples*, which does not.

Appendix A.4 points out that verbs sometimes change their interpretation if they occur in a context that calls for items of a different aspectual class than their own. In order to avoid a clash, the interpretation of the verb is changed without visible specification. This 'coercion' or 'reinterpretation' can complicate the annotation.

Annotation takes place with the help of an annotation tool that was specifically developed for this task and is described in appendix A.5.

A.2 ASPECTUAL CLASSES

This section introduces the aspectual dichotomies according to which the aspectual classification takes place. After a short characterisation of the dichotomy, aspectual 'tests' will be provided, which help you to make the distinction for a given verb. Such tests take the form of linguistic contexts in which items of a specific aspectual class can or cannot occur. For a list of aspectual tests, see Dowty (1979). Note that these tests are only valid if the meaning of the verb is not coerced in order to fit in with an aspectual restriction of the environment, see appendix A.4 for details.

Some of the examples used in this section consist of a verb with an argument. These are cases in which a specific argument exerts an aspectual influence, see appendix A.3 for a detailed account.

A.2.1 *Valid - Invalid*

This first distinction is not a true aspectual distinction, but since it is part of the annotation, it is grouped here with the aspectual classification.

Whenever you annotate a new sentence, the first choice pertains to the question of whether you are dealing with an item that can reasonably be classified aspectually. Because the pre-processing of the annotation corpus has been done automatically, the corpus may contain items that are not appropriate to the primary task. In these cases, the verb should be tagged as invalid. Examples include:

- Mis-tagged verbs, in particular, auxiliaries (which are excluded from the annotation task, but might show up nevertheless in the annotation task because they have been falsely classified as a main verb), or verbs with separable prefixes where the prefix has not been recognised as being part of the verb);
- Sentence fragments, where complements to a verb needed for the assessment of lexical aspect are not shown (see appendix A.3), and it is not even possible to define the range of possibilities;

- Fragmentary expressions in which the exact meaning of the verb cannot be ascertained, including cases in which the verb itself is missing.

A.2.2 *Stative - Dynamic*

Stative verbs describe a static situation without any change, e.g., *mögen* 'like', whereas dynamic verbs introduce eventualities with some change (e.g., continuous change of position in *move*). Stative verbs express that a property obtains at a particular point or interval in time. They do not refer to eventualities with inherent boundaries, or to eventualities like processes or events that can 'happen'.

Stative verbs are frequently incompatible with the imperative mood.

- *Geh nach Hause!* 'Go home!'
- ? *Wisse die Antwort!* 'Know the answer!'

Stative verbs cannot be the complement of the verb *zwingen* 'to force' or occur in combination with adverbials expressing intentionality like *freiwillig* 'voluntarily':

- *Ich habe sie gezwungen, nach Hause zu gehen.* 'I forced her to go home.'
- ? *Ich habe sie gezwungen, die Antwort zu wissen.* 'I forced her to know the answer.'
- ? *Er weiß die Antwort freiwillig.* 'He knows the answer voluntarily.'
- *Er geht freiwillig nach Hause.* 'He is going home voluntarily.'

Also, if the validity of a verb for a specific temporal interval entails its validity for every instant within this interval, we are dealing with a stative verb, e.g. *beam*. This is different for dynamic verbs like *walk*, so, if John beamed from 8-9am, he also beamed at 8.30pm sharp, but if he walked from 8-9am, we cannot claim that he walked at 8.30 sharp, because walking eventualities have a certain minimal size (1-2 steps).

Note that one of the most prominent tests to distinguish dynamic from stative verbs, viz., the progressive, is not available for German.

A.2.3 *Unbounded - Bounded*

Dynamic verbs can be unbounded, i.e., introduce eventualities without inherent boundaries, e.g., *move* or *play the piano*, or bounded, i.e., refer to eventualities with such boundaries, e.g., *run a mile* or *build a house*. Eventualities with boundaries have a natural conclusion, a point in

time at which the expressed action is finished and cannot continue any longer. Eventualities without these boundaries cannot be finished in this way, they can only be stopped.

Boundedness of a verb can usually be determined by combining it with a time frame adverbial like *in zwei Stunden* 'in two hours' that express the (maximal) amount of time required to finish the action. Consequently, unbounded items like *Klavier spielen* 'play the piano' are not compatible with these adverbials:

- *? Ich habe in fünfzehn Minuten Klavier gespielt.* 'I played the piano in 15 minutes.'
- *Ich habe in fünfzehn Minuten die Sonate gespielt.* 'I played the sonata in 15 minutes.'

The combination with durative adverbials like *zwei Stunden (lang)* 'for two hours' has exactly the opposite effect. Since such adverbials introduce boundaries themselves, they are only compatible with unbounded, but not with bounded items:

- *Ich habe fünfzehn Minuten (lang) Klavier gespielt.* 'I played the piano for 15 minutes.'
- *? Ich habe fünfzehn Minuten (lang) die Sonate gespielt.* 'I played the sonata for 15 minutes.'

A.2.4 *Punctual - Durative*

Events may be considered to take a particular amount of time, or they may be conceived of as happening (relatively) instantaneously. To distinguish punctual and durative bounded verbs, we can use adverbials like *plötzlich* 'suddenly' or time point PPs like *um 11 Uhr* 'at 11 o'clock', which are only compatible with the first group:

- *Er hat plötzlich/um 11 Uhr gehustet.* 'He coughed suddenly/at 11 o'clock.'
- *? Er hat plötzlich/um 11 Uhr ein Haus gebaut.* 'He built a house suddenly/at 11 o'clock.'

Similarly, time frame adverbials can only measure the duration of eventualities denoted by durative verbs like *ein Haus bauen* 'build a house' but not of punctual verbs like *husten* 'cough'.

- *Er baute in drei Monaten ein Haus.* 'He built a house in three months.'
- *? Er hustete in einer Sekunde.* 'He coughed in one second.'

A.2.5 *Change - No change*

Change verbs lexically express a specific change of state in their semantics, e.g., *sterben* 'die' expresses a change from being alive to being dead. I.e., they characterise both the situation immediately before the eventualities they refer to (the 'prestate') as well as the situation immediately afterwards (the 'poststate').

This is in contrast with bounded verbs that introduce an eventuality with inherent boundaries but no explicit change of state, e.g., *husten* 'cough' or *eine Meile laufen* 'run a mile'. They merely indicate that a specific activity lasted for a certain amount of time (extended or punctual), without characterising pre- and poststate.

Verbs of change (especially those denoting an externally caused change) often permit an alternation between the causative and the inchoative:

- *Das Fenster zerbrach.* 'The window broke.'
- *Johann zerbrach das Fenster.* 'Johann broke the window.'

Changes of state may involve the change of location or possession of a patient argument.

- *Ich schenkte meiner Mutter die Rosen.* 'I gave my mother the roses.'
- *Sie holte einen Rechen aus dem Geräteschuppen.* 'I fetched a rake from the garden shed.'

The change introduced by the verb may be temporary or permanent, and may vary between borderline changes of state which alter an entity superficially to events which result in the creation or destruction of an entity.

- *Ich errötete.* 'I blushed.'
- *Ich wischte mir die Stirn ab.* 'I wiped my forehead.'
- *Ich öffnete die Tür.* 'I opened the door.'
- *Ich habe das Buch geschrieben/den Kuchen gebacken.* 'I wrote the book / baked the cake.'
- *Ich habe den Brief verbrannt.* 'I burned the letter.'

The standard test for change verbs in English is the perfect tense, however, since the perfect in German is gradually taking over the role of the imperfective, it is no longer a reliable test for aspectual class in German. Still, we can simulate the effect of this test by combining the perfect with *gerade eben* 'just now':

- *Max ist gerade eben angekommen.* 'Max has arrived just now.'

- *Max ist gerade eben gelaufen.* 'Max has run just now.'
- *Max hat gerade eben gehustet.* 'Max has coughed just now.'

If such a sentence entails information on the moment of utterance, we are dealing with a change verb. While it is possible in German to combine the perfect plus *gerade eben* with no-change verbs like *husten* 'cough' or *laufen* 'run', too, this does not entail anything about the moment of utterance. This is in contrast to the first example, which entails that at the moment of utterance, Max is present, because *ankommen* 'arrive' introduces a change of state from being away to being present.

Another aspectual test for change verbs uses *fast* 'almost'. The combination of change verbs with *fast* leads to ambiguity, e.g., for *fast sterben* 'almost die':

- *John starb fast* 'John almost died.'

In the first reading, John came close to undergoing a change from being alive to being dead, but, in fact, nothing happened; in the second reading, he did undergo a change, from being alive to being close to dead. No such ambiguity would arise for *fast husten* 'almost cough', which only has the reading that one came close to coughing, but that, eventually, nothing happened.

A.3 MARKING VERBS IN CONTEXT

In the preceding section, the examples sometimes involved not only verbs, but an additional argument. The reason for this is that in some cases, the verb in isolation is not sufficient for aspectual classification.

The temporal progression of an eventuality is primarily characterised by a verb. While in many cases the verb already determines this progression, and hence can be classified aspectually in isolation, specific arguments of a verb can also influence the aspectual class, especially if it refers to an entity that is affected or changed by the eventuality introduced by the verb gradually. If so, this argument is important for the verb's aspectual classification and must be taken into consideration for the annotation.

In particular, this holds for so-called 'incremental themes' in verbs like *essen* 'eat' or *bauen* 'build' (Krifka, 1992). The idea is that the object referents undergo the process of eating or building gradually, hence, any boundaries that the objects introduce carry over to the eventuality itself: For instance, eating an apple will inevitably reach the stage at which the last bit of the apple has been consumed, at which moment the eventuality necessarily stops. The resulting classification for *einen Apfel essen* 'eat an apple' is extended and change.

No such boundaries are introduced by *Äpfel essen* 'eat apples', because the object *Äpfel* itself introduces no such boundaries. Hence, it classifies as dynamic and unbounded.

Note that such incremental themes can be of a more abstract nature, which shows up for *eine Sonate spielen* 'play a sonata' vs. *Sonaten spielen* 'play sonatas' as well as for *eine Meile laufen* 'walk a mile' vs. *ein paar Meter laufen* 'walk a few metres'. Here the pieces of music and the traversed paths are the gradually processed entities. Again, if they introduce boundaries, these are inherited by the verb as well. The resulting class in these cases would be extended no-change, because neither *eine Sonate spielen* nor *eine Meile laufen* introduce a change of state.

In combination with objects that introduce no boundaries like in *Sonaten spielen* and *ein paar Meter laufen*, the verbs qualify as dynamic and unbounded.

At this point, it is necessary to distinguish this effect, which is obligatory, from similar phenomena, which are only optional, like in the following example.

- *Stundenlang kamen Gäste an.* 'For hours, guests arrived.'

Although we have classified *ankommen* 'arrive' as a punctual change verb, which due to its boundedness should not be compatible with durative adverbials like *stundenlang* 'for hours', its argument seems to exert the same kind of influence that we have seen for incremental arguments: If the guests arrived one after the other, and the number is not limited, this unboundedness seems to be inherited by the verb, which then licenses compatibility with the durative adverbial.

However, the marked difference between this case and true gradual arguments is its optionality. Compare:

- *Plötzlich kamen Gäste an.* 'Suddenly, guests arrived.'
- *? Plötzlich aß er einen Apfel.* 'Suddenly, he ate an apple.'

The first sentence is fully acceptable (in the sense that the guests arrived as one group), while the second one can only be understood very marginally (if at all) in a non-literal sense (that the beginning of the eating was immediate and unexpected). Such non-literal interpretations, however, are due to the additional process of coercion (see the next section), hence, show that an aspectual selection restriction is violated. In other words, the verb in the first sentence is punctual, because it can be combined with *plötzlich*, while the second verb is not. The fact that *ankommen* in combination with a bare plural can be understood in an unbounded way is due to a secondary optional process of iteration.

The annotation should only include the effect of verb arguments if they are mandatory like in the case of *essen* or *spielen*, and ignore optionally influencing arguments like for *ankommen*.

A.4 ASPECTUAL COERCION OR REINTERPRETATION

Sometimes a verb is 'coerced' or 'reinterpreted' to fit aspectual requirements of the operator. E.g., the English progressive requires nonstative and extended predicates, still the stative *be silly* and the punctual *cough* can occur in the progressive, after being coerced into unbounded predicates ('act in a silly way' and 'cough repeatedly', respectively). For a list of such coercion patterns, see Moens and Steedman (1988).

A typical kind of coercion applies to punctual no-change verbs like *husten* 'cough' or *klopfen* 'knock': If they are combined with a durative adverbial (which requires an unbounded verb), an iterative operator can intervene, which changes the interpretation of the verb into 'performing a whole series of eventualities of the kind denoted by the literal meaning of the verb'. Such an iteration is unbounded and hence compatible with the adverbial.

Similarly, the combination of extended verbs with adverbials like *plötzlich* 'suddenly' that require a punctual verb can cause inchoative reinterpretation of the verb (because the beginning of an extended eventuality is only a moment of time). For instance, *plötzlich rennen* 'suddenly run' refers to the start of a running that happens quickly and unexpectedly.

The effect of such coercions should not be reflected in the annotation, which strives at recording the original aspectual class, the one that obtained before reinterpretation took place.

A.5 THE ANNOTATION TOOL

The tool is to be found online under `https://www.annotate.wkroberts.com`. When you start the annotation work, you need to register and create a new account. Each sentence is presented for annotation on a page of its own. You can go backward and forward, e.g., in order to compare or change previous annotations. The database is searchable for individual verbs or text in general. Note that you first indicate the material to be searched, then you can move to sentences comprising this material by clicking 'Previous clause' or 'Next clause'.

You will find that the choices that you make are interdependent, for instance, if you classify a verb as unbounded, it will automatically also be classified as dynamic. These dependencies are motivated by the underlying aspectual theory and safeguard that erroneous combinations of aspectual features are ruled out automatically (e.g., stative and bounded would not be possible to annotate).

The tool allows you to model aspectual ambiguity in two ways. First, you can assign several full aspectual annotations by clicking the 'New annotation' button at the bottom of the page. This is a phenomenon that can be quite systematic, e.g., many verbs in the

semantic field of communication like *zeigen* 'show' or *beweisen* 'prove' have a stative reading, here, 'be a logical implicature for' and simultaneously an extended change reading, here, 'perform a logical deduction'. Similarly, so-called 'degree achievements' like *den Weg kehren* 'sweep the path' have both an unbounded reading (continuous development, here, towards cleanliness) and an extended change reading (here, crossing a threshold of cleanliness).

Second, in between the two elements in a feature pair there is a question mark, which you can use in case you feel unable to select one of them.

INDEX OF TERMS

BIBLIOGRAPHY

Abney, Steven P. (1997). 'Stochastic attribute-value grammars'. In: *Computational Linguistics* 23.4, pp. 597–618.

Alikhani, Malihe and Matthew Stone (2019). '"Caption" as a coherence relation: Evidence and implications'. In: *Proceedings of the Second Workshop on Shortcomings in Vision and Language*. Minneapolis, MN: Association for Computational Linguistics, pp. 58–67.

Aminian, Maryam, Mohammad Sadegh Rasooli and Hossein Sameti (2013). 'Unsupervised induction of Persian semantic verb classes based on syntactic information'. In: *Language Processing and Intelligent Information Systems*. Ed. by Mieczysław A. Kłopotek, Jacek Koronacki, Małgorzata Marciniak, Agnieszka Mykowiecka and Sławomir T. Wierzchoń. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 112–124.

Andrews, Avery (1985). 'The major functions of the noun phrase'. In: *Language typology and syntactic description, Vol. I: Clause structure*. Ed. by Timothy Shopen. Cambridge: Cambridge University Press. Chap. 2, pp. 62–154.

Aparicio, Juan, Mariona Taulé and Maria Antònia Martí (2008). 'AnCora-Verb: A lexical resource for the semantic annotation of corpora'. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Attardi, Giuseppe and Felice Dell'Orletta (2009). 'Reverse revision and linear tree combination for dependency parsing'. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Boulder, CO: Association for Computational Linguistics, pp. 261–264.

Bach, Emmon (1981). 'Time, tense, and aspect: An essay in English metaphysics'. In: *Radical pragmatics*. Ed. by Peter Cole. New York, NY: Academic Press, pp. 63–81.

— (1986). 'The algebra of events'. In: *Linguistics and Philosophy* 9.1, pp. 5–16.

Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio (2015). 'Neural machine translation by jointly learning to align and translate'. In: *Proceedings of the International Conference on Learning Representations (ICLR 2015)*. San Diego, CA. arXiv: 1409.0473.

Baiamonte, Daniela, Tommaso Caselli and Irina Prodanof (2016). 'Annotating content zones in news articles'. In: *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016)*.

Baker, Collin F., Charles J. Fillmore and John B. Lowe (1998). 'The Berkeley FrameNet project'. In: *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics*. Montréal, Canada: Association for Computational Linguistics, pp. 86–90.

Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston and Marco Mazzoleni (2004). 'Introducing the La Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian'. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA).

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta (2009). 'The WaCky wide web: A collection of very large linguistically processed Web-crawled corpora'. In: *Language Resources and Evaluation* 43.3, pp. 209–226.

Baroni, Marco, 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)*. Baltimore, MD: Association for Computational Linguistics, pp. 238–247.

Bauer, Laurie and Salvador Valera Hernández (2005). 'Conversion or zero-derivation: An introduction'. In: *Approaches to conversion / Zero-derivation*. Ed. by Laurie Bauer and Salvador Valera Hernández. Münster, Germany: Waxmann Verlag GmbH. Chap. 1, pp. 7–18.

Baum, Leonard E. (1972). 'An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes'. In: *Inequalities* 3, pp. 1–8.

Berger, Adam L., Stephen A. Della Pietra and Vincent J. Della Pietra (1996). 'A maximum entropy approach to natural language processing'. In: *Computational Linguistics* 22.1, pp. 39–71.

Bergsma, Shane, Dekang Lin and Randy Goebel (2008). 'Discriminative learning of selectional preference from unlabeled text'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, HI: Association for Computational Linguistics, pp. 59–68.

Bikel, Daniel M. (2004). 'On the parameter space of generative lexicalized statistical parsing models'. PhD thesis. University of Pennsylvania.

Bird, Steven, Ewan Klein and Edward Loper (2009). 'Natural language processing with Python'. Sebastopol, CA: O'Reilly Media.

Bittar, André, Pascal Amsili, Pascal Denis and Laurence Danlos (2011). 'French TimeBank: An ISO-TimeML annotated reference corpus'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR: Association for Computational Linguistics, pp. 130–134.

Björkelund, Anders, Love Hafdell and Pierre Nugues (2009). 'Multilingual semantic role labeling'. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*. Boulder, CO: Association for Computational Linguistics, pp. 43–48.

Black, Ezra et al. (1991). 'A procedure for quantitatively comparing the syntactic coverage of English grammars'. In: *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Blei, David M., Andrew Y. Ng and Michael I. Jordan (2003). 'Latent Dirichlet allocation'. In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Bloomfield, Leonard (1933). 'Language'. New York, NY: Holt.

Boguraev, Bran and Ted Briscoe (1987). 'Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE'. In: *Computational Linguistics* 13.3–4, pp. 203–218.

Boguraev, Bran, Ted Briscoe, John Carroll, David Carter and Claire Grover (1987). 'The derivation of a grammatically indexed lexicon from the Longman dictionary of contemporary English'. In: *25th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 193–200.

Bohnet, Bernd (2010). 'Top accuracy and fast dependency parsing is not a contradiction'. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, pp. 89–97.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter and Jan Hajič (2013). 'Joint morphological and syntactic analysis for richly inflected languages'. In: *Transactions of the Association for Computational Linguistics* 1, pp. 415–428.

Bojar, Ondřej et al. (2016). 'Findings of the 2016 conference on machine translation'. In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, pp. 131–198.

Box, George E. P. and David R. Cox (1964). 'An analysis of transformations'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2, pp. 211–252.

Boyd-Graber, Jordan, David Blei and Xiaojin Zhu (2007). 'A topic model for word sense disambiguation'. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 1024–1033.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith (2002). 'The TIGER treebank'. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pp. 24–41.

Breiman, Leo (2001). 'Random forests'. In: *Machine Learning* 45.1, pp. 5–32.

Brent, Michael R. (1991). 'Automatic acquisition of subcategorization frames from untagged text'. In: *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*. Berkeley, CA: Association for Computational Linguistics, pp. 209–214.

— (1993). 'From grammar to lexicon: Unsupervised learning of lexical syntax'. In: *Computational Linguistics* 19.2, pp. 243–262.

Bresnan, Joan (2001). 'Lexical-functional syntax'. Malden, MA: Blackwell.

Brew, Chris and Sabine Schulte im Walde (2002). 'Spectral clustering for German verbs'. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 117–124.

Briscoe, Ted and John Carroll (1997). 'Automatic extraction of subcategorization from corpora'. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics. Washington, DC, pp. 356–363.

Brockmann, Carsten and Mirella Lapata (2003). 'Evaluating and combining approaches to selectional preference acquisition'. In: *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*. Budapest, Hungary, pp. 27–34.

Brown, Peter F., Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra and Jenifer C. Lai (1992). 'Class-based n-gram models of natural language'. In: *Computational Linguistics* 18.4, pp. 467–479.

Brown, Tom B. et al. (2020). 'Language models are few-shot learners'. In: arXiv: `2005.14165`.

Buchholz, Sabine and Erwin Marsi (2006). 'CoNLL-X shared task on multilingual dependency parsing'. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York, NY: Association for Computational Linguistics, pp. 149–164.

Bühler, Karl (1982). 'The deictic field of language and deictic words'. In: *Speech, place and action*. Ed. by Robert J. Jarvella and Wolfgang Klein. Chichester: John Wiley & Sons, pp. 9–30.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó and Manfred Pinkal (2006). 'The SALSA corpus: A German corpus resource for lexical semantics'. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Burnard, Lou (2000). *The British National Corpus Users Reference Guide*. Tech. rep.

Carroll, John and Alex C. Fang (2004). 'The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser'. In: *Natural Language Processing - IJCNLP 2004, First International Joint Conference*. Ed. by Keh-Yih Su, Jun'ichi Tsujii,

Jong-Hyeok Lee and Oi Yee Kwong. Vol. 3248. Lecture Notes in Computer Science. Hainan Island, China: Springer, pp. 646–654.

Carroll, John, Guido Minnen and Ted Briscoe (1998). 'Can subcategorisation probabilities help a statistical parser'. In: *Sixth Workshop on Very Large Corpora*.

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta and Irina Prodanof (2011). 'Annotating events, temporal expressions and relations in Italian: The It-TimeML experience for the Ita-TimeBank'. In: *Proceedings of the 5th Linguistic Annotation Workshop*. Portland, OR: Association for Computational Linguistics, pp. 143–151.

Caselli, Tommaso and Valeria Quochi (2007). 'Inferring the semantics of temporal prepositions in Italian'. In: *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*. Prague, Czech Republic: Association for Computational Linguistics, pp. 38–44.

Chafe, Wallace L. (1970). 'Meaning and the structure of language'. Chicago, IL: University of Chicago Press.

Chambers, Nathanael and Daniel Jurafsky (2010). 'Improving the use of pseudo-words for evaluating selectional preferences'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 445–453.

Charniak, Eugene (1997). 'Statistical parsing with a context-free grammar and word statistics'. In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 598–603.

— (2000). 'A maximum-entropy-inspired parser'. In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 132–139.

Charniak, Eugene and Mark Johnson (2005). 'Coarse-to-fine n-best parsing and MaxEnt discriminative reranking'. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, MI: Association for Computational Linguistics, pp. 173–180.

Chen, Danqi and Christopher D. Manning (2014). 'A fast and accurate dependency parser using neural networks'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 740–750.

Chen, Tianqi and Carlos Guestrin (2016). 'XGBoost: A scalable tree boosting system'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 785–794.

Chesley, Paula and Susanne Salmon-Alt (2006). 'Automatic extraction of subcategorization frames for French'. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*

(LREC'06). Genoa, Italy: European Language Resources Association (ELRA).

Chomsky, Noam (1957). 'Syntactic structures'. The Hague: Mouton.

— (1981). 'Lectures on government and binding'. Dordrecht: Foris.

Chung, Sandra and Alan Timberlake (1985). 'Tense, aspect and mood'. In: *Language typology and syntactic description, Vol. 3: Grammatical categories and the lexicon*. Ed. by Timothy Shopen. Cambridge: Cambridge University Press. Chap. 5, pp. 202–258.

Clark, Stephen and David Weir (2001). 'Class-based probability estimation using a semantic hierarchy'. In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pp. 1–8.

— (2002). 'Class-based probability estimation using a semantic hierarchy'. In: *Computational Linguistics* 28.2, pp. 187–206.

Cliche, Mathieu (2017). 'BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs'. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 573–580.

Collins, Michael (2003). 'Head-driven statistical models for natural language parsing'. In: *Computational Linguistics* 29.4, pp. 589–637.

Collobert, Ronan and Jason Weston (2008). 'A unified architecture for natural language processing: Deep neural networks with multitask learning'. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuksa (2011). 'Natural language processing (almost) from scratch'. In: *Journal of Machine Learning Research* 12, pp. 2493–2537.

Comrie, Bernard (1976). 'Aspect'. Cambridge: Cambridge University Press.

— (1985). 'Tense'. Cambridge: Cambridge University Press.

Costa, Francisco and António Branco (2012a). 'Aspectual type and temporal relation classification'. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 266–275.

— (2012b). 'TimeBankPT: A TimeML annotated corpus of Portuguese'. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3727–3734.

Croft, William (2012). 'Verbs: Aspect and causal structure'. Oxford: Oxford University Press.

Croft, William, Pavlina Peskova and Michael Regan (2016). 'Annotation of causal and aspectual structure of events in RED: A prelim-

inary report'. In: *Proceedings of the Fourth Workshop on Events*. San Diego, CA: Association for Computational Linguistics, pp. 8–17.

Cruse, D. Alan (1986). 'Lexical semantics'. Cambridge: Cambridge University Press.

Culicover, Peter W. and Wendy Wilkins (1986). 'Control, PRO, and the projection principle'. In: *Language* 62.1, pp. 120–153.

Davidson, Donald (1967). 'The logical form of action sentences'. In: *The logic of decision and action*. Ed. by Nicholas Rescher. Pittsburgh, PA: University of Pittsburgh Press.

Davis, Anthony R. (2011). 'Thematic roles'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Mouton de Gruyter. Chap. 18, pp. 399–420.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). 'BERT: Pre-training of deep bidirectional transformers for language understanding'. In: arXiv: 1810.04805.

Dipper, Stefanie and Sandra Kübler (2017). 'German treebanks: TIGER and TüBa-D/Z'. In: *Handbook of Linguistic Annotation*. Ed. by Nancy Ide and James Pustejovsky. Berlin: Springer, pp. 595–639.

Dorr, Bonnie J. (1997). 'Large-scale dictionary construction for foreign language tutoring and interlingual machine translation'. In: *Machine Translation* 12.4, pp. 271–322.

Dorr, Bonnie J. and Doug Jones (1996). 'Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues'. In: *Proceedings of the 16th Conference on Computational Linguistics*. Vol. 1, pp. 322–327.

Dorr, Bonnie J., Mari Broman Olsen, Nizar Habash and Scott Thomas (2001). 'LCS verb database'. In: *Online Software Database of Lexical Conceptual Structures*.

Dowty, David R. (1979). 'Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ'. Dordrecht: Reidel.

— (1991). 'Thematic proto-roles and argument selection'. In: *Language* 67.3, pp. 547–619.

Dubois, Jean and Françoise Dubois-Charlier (1997). 'Les Verbes français'. Larousse.

Dudenredaktion, ed. (2001). 'Duden - Das Stilwörterbuch'. 8th ed. Mannheim: Dudenverlag.

Dunning, Ted (1993). 'Accurate methods for the statistics of surprise and coincidence'. In: *Computational Linguistics* 19.1, pp. 61–74.

Dyer, Chris, Miguel Ballesteros, Wang Ling, Austin Matthews and Noah A. Smith (2015). 'Transition-based dependency parsing with stack long short-term memory'. 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). Beijing, China: Association for Computational Linguistics, pp. 334–343.

Eckle-Kohler, Judith (1999). 'Linguistic knowledge for automatic lexicon acquisition from German text corpora'. PhD thesis. Universität Stuttgart.

Egg, Markus (1995). 'The intergressive as a new category of verbal Aktionsart'. In: *Journal of Semantics* 12.4, pp. 311–356.

— (2005). 'Flexible semantics for reinterpretation phenomena'. Stanford, CA: Center for the Study of Language (CSLI).

Egg, Markus, Helena Prepens and Will Roberts (2019). 'Annotation and automatic classification of aspectual categories'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3335–3341.

Engelberg, Stefan (2011a). 'Frameworks of lexical decomposition of verbs'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Vol. 1. Berlin: Mouton de Gruyter. Chap. 17, pp. 358–399.

— (2011b). 'Lexical decomposition: Foundational issues'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Vol. 1. Berlin: Mouton de Gruyter. Chap. 7, pp. 124–144.

Erk, Katrin (2007). 'A simple, similarity-based model for selectional preferences'. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, pp. 216–223.

Erk, Katrin, Sebastian Padó and Ulrike Padó (2010). 'A flexible, corpus-driven model of regular and inverse selectional preferences'. In: *Computational Linguistics* 36.4, pp. 723–763.

Esteve Ferrer, Eva (2004). 'Towards a semantic classification of Spanish verbs based on subcategorisation information'. In: *Proceedings of the ACL Student Research Workshop*. Barcelona, Spain: Association for Computational Linguistics, pp. 37–42.

Faaß, Gertrud and Kerstin Eckart (2013). 'SdeWaC - A corpus of parsable sentences from the Web'. In: *Language processing and knowledge in the Web*. Berlin, Heidelberg: Springer, pp. 61–68.

Falk, Ingrid, Claire Gardent and Jean-Charles Lamirel (2012). 'Classifying French verbs using French and English lexical resources'. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 854–863.

Falk, Ingrid and Fabienne Martin (2016). 'Automatic identification of aspectual classes across verbal readings'. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Berlin, Germany: Association for Computational Linguistics, pp. 12–22.

Fellbaum, Christiane (1990). 'English verbs as a semantic net'. In: *International Journal of Lexicography* 3.4, pp. 278–301.

— ed. (1998). 'WordNet: An electronic lexical database'. Cambridge, MA: MIT Press.

Filip, Hana (2012). 'Aspectual class and aktionsart'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Vol. 2. Berlin: Mouton de Gruyter. Chap. 48, pp. 1186–1217.

Fillmore, Charles J. (1976). 'Frame semantics and the nature of language'. In: *Annals of the New York Academy of Sciences* 280, pp. 20–32.

— (2003). 'The grammar of hitting and breaking'. In: *Form and meaning in language*. CSLI Publications, Center for the Study of Language and Information. Chap. 3, pp. 123–139.

Firth, John R. (1957). 'A synopsis of linguistic theory 1930–1955'. In: *Studies in linguistic analysis*. Oxford: Blackwell, pp. 1–32.

Forgy, Edward W. (1965). 'Cluster analysis of multivariate data: Efficiency versus interpretability of classifications'. In: *Biometrics* 21, pp. 768–769.

Frawley, William (1992). 'Linguistic semantics'. Hillsdale, NJ: Lawrence Erlbaum.

Friedrich, Annemarie and Alexis Palmer (2014). 'Automatic prediction of aspectual class of verbs in context'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, MD: Association for Computational Linguistics, pp. 517–523.

Friedrich, Annemarie, Alexis Palmer and Manfred Pinkal (2016). 'Situation entity types: Automatic classification of clause-level aspect'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1757–1768.

Friedrich, Annemarie, Alexis Palmer, Melissa Peate Sørensen, Manfred Pinkal, Melissa Peate Sørensen and Manfred Pinkal (2015). 'Annotating genericity: A survey, a scheme, and a corpus'. In: *The 9th Linguistic Annotation Workshop*. Denver, CO: Association for Computational Linguistics, pp. 21–30.

Friedrich, Annemarie and Manfred Pinkal (2015). 'Automatic recognition of habituals: A three-way classification of clausal aspect'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2471–2481.

Gildea, Daniel and Daniel Jurafsky (2002). 'Automatic labeling of semantic roles'. In: *Computational Linguistics* 28.3, pp. 245–288.

Givón, Talmy (1979). 'On understanding grammar'. Academic Press.

Govindarajan, Venkata, Benjamin Van Durme and Aaron Steven White (2019). 'Decomposing generalization: Models of generic, habitual,

and episodic statements'. In: *Transactions of the Association for Computational Linguistics* 7, pp. 501–517.

Graff, David, Junbo Kong, Ke Chen and Kazuaki Maeda (2003). *English gigaword*. Tech. rep.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin and Tomas Mikolov (2018). 'Learning word vectors for 157 languages'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Languages Resources Association (ELRA).

Grenager, Trond and Christopher D. Manning (2006). 'Unsupervised discovery of a statistical verb lexicon'. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia: Association for Computational Linguistics, pp. 1–8.

Grishman, Ralph, Catherine Macleod and Adam Meyers (1994). 'Comlex syntax: Building a computational lexicon'. In: *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Guo, Yufan, Anna Korhonen and Thierry Poibeau (2011). 'A weakly-supervised approach to argumentative zoning of scientific documents'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK: Association for Computational Linguistics, pp. 273–283.

Gurevych, Iryna and Hendrik Niederlich (2005). 'Computing semantic relatedness in German with revised information content metrics'. In: *Proceedings of "OntoLex 2005 - Ontologies and Lexical Resources" IJCNLP'05 Workshop*, pp. 28–33.

Habash, Nizar, Bonnie J. Dorr and David R. Traum (2003). 'Hybrid natural language generation from lexical conceptual structures'. In: *Machine Translation* 18.2, pp. 81–128.

Hajič, Jan and Barbora Hladká (1998). 'Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset'. In: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Hajič, Jan et al. (2009). 'The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages'. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*. Boulder, CO: Association for Computational Linguistics, pp. 1–18.

Hall, Johan and Joakim Nivre (2008). 'A dependency-driven parser for German dependency and constituency representations'. In: *Proceedings of the Workshop on Parsing German*. Columbus, OH: Association for Computational Linguistics, pp. 47–54.

Hammer, Alfred Edward (1983). 'German grammar and usage'. London: Edward Arnold.

Hamp, Birgit and Helmut Feldweg (1997). 'GermaNet: A lexical-semantic net for German'. In: *Proceedings of ACL Workshop on Automatic In-*

formation Extraction and Building of Lexical Semantic Resources for NLP Applications, pp. 9–15.

Harris, Zellig S. (1954). 'Distributional structure'. In: *Word* 10.23, pp. 146–162.

Hartmann, Silvana, Éva Mújdricza-Maydt, Ilia Kuznetsov, Iryna Gurevych and Anette Frank (2017). 'Assessing SRL frameworks with automatic training data expansion'. In: *Proceedings of the 11th Linguistic Annotation Workshop*. Valencia, Spain: Association for Computational Linguistics, pp. 115–121.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown (1993). 'Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning'. In: *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*. Columbus, OH, pp. 172–182.

Herweg, Michael (1991). 'Temporale Konjunktionen und Aspekt. Der sprachliche Ausdruck von Zeitrelationen zwischen Situationen'. In: *Kognitionswissenschaft* 2.2, pp. 51–90.

Hornby, Albert Sydney (1974). 'Oxford Advanced Learner's Dictionary of Current English'. London: Oxford University Press.

Hubert, Lawrence and Phipps Arabie (1985). 'Comparing partitions'. In: *Journal of Classification* 2.1, pp. 193–218.

Ide, Nancy, Collin F. Baker, Christiane Fellbaum and Rebecca J. Passonneau (2010). 'The Manually Annotated Sub-Corpus: A community resource for and by the people'. In: *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, pp. 68–73.

Ikuta, Rei, Will Styler, Mariah Hamang, Tim O'Gorman and Martha Palmer (2014). 'Challenges of adding causation to richer event descriptions'. In: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Baltimore, MD: Association for Computational Linguistics, pp. 12–20.

Iordachioaia, Gianina, Lonneke van der Plas and Glorianna Jagfeld (2016). 'The grammar of English deverbal compounds and their meaning'. In: *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 81–91.

Iyyer, Mohit, Varun Manjunatha, Jordan Boyd-Graber and Hal Daumé III (2015). 'Deep unordered composition rivals syntactic methods for text classification'. 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)*. Beijing, China: Association for Computational Linguistics, pp. 1681–1691.

Jackendoff, Ray (1983). 'Semantics and cognition'. Cambridge, MA: MIT Press.

Jackendoff, Ray (1987). 'Consciousness and the computational mind'. Cambridge, MA: MIT Press.

Joanis, Eric (2002). 'Automatic verb classification using a general feature space'. Master's thesis. University of Toronto.

Joanis, Eric, Suzanne Stevenson and David James (2008). 'A general feature space for automatic verb classification'. In: *Natural Language Engineering* 14.3, pp. 337–367.

Kamholz, David, Jonathan Pool and Susan M. Colowick (2014). 'PanLex: Building a resource for panlingual lexical translation'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3145–3150.

Kaplan, Ronald M. and Joan Bresnan (1982). 'Lexical-functional grammar: A formal system for grammatical representation'. In: *The mental representation of grammatical relations*. Ed. by Joan Bresnan. Cambridge, MA: MIT Press, pp. 173–281.

Katz, Jerrold J. and Jerry A. Fodor (1963). 'The structure of a semantic theory'. In: *Language* 39.2, pp. 170–210.

Katz, Slava (1987). 'Estimation of probabilities from sparse data for the language model component of a speech recognizer'. In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 35.3, pp. 400–401.

Kawahara, Daisuke, Daniel W. Peterson and Martha Palmer (2014). 'A step-wise usage-based method for inducing polysemy-aware verb classes'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, MD, pp. 1030–1040.

Keller, Frank and Mirella Lapata (2003). 'Using the Web to obtain frequencies for unseen bigrams'. In: *Computational Linguistics* 29.3, pp. 459–484.

Kennedy, Christopher (2011). 'Ambiguity and vagueness: An overview'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Vol. 1. Berlin: Mouton de Gruyter. Chap. 23, pp. 507–535.

Kennedy, Christopher and Beth Levin (2008). 'Measure of change: The adjectival core of degree achievements'. In: *Adjectives and adverbs: Syntax, semantics and discourse*. Ed. by Louise McNally and Chris Kennedy. Oxford: Oxford University Press, pp. 156–182.

Kenny, Anthony (1963). 'Action, emotion and will'. London, New York: Routledge.

Kilgarriff, Adam and Gregory Grefenstette (2003). 'Introduction to the special issue on the Web as corpus'. In: *Computational Linguistics* 29.3, pp. 333–348.

Kim, Yoon (2014). 'Convolutional neural networks for sentence classification'. In: *Proceedings of the 2014 Conference on Empirical Methods*

in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, pp. 1746–1751.

Kipper-Schuler, Karin (2005). 'VerbNet: A broad-coverage, comprehensive verb lexicon'. PhD thesis. University of Pennsylvania.

Kipper, Karin, Anna Korhonen, Neville Ryant and Martha Palmer (2006). 'Extending VerbNet with novel verb classes'. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Klavans, Judith L. and Martin Chodorow (1992). 'Degrees of stativity: The lexical representation of verb aspect'. In: *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*, pp. 1126–1131.

Klavans, Judith L. and Min-Yen Kan (1998). 'Role of verbs in document analysis'. In: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Klein, Dan and Christopher D. Manning (2003). 'Accurate unlexicalized parsing'. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Sapporo, Japan, pp. 423–430.

Klein, Wolfgang (1994). 'Time in language'. London: Routledge.

Kohomban, Upali Sathyajith and Wee Sun Lee (2005). 'Learning semantic classes for word sense disambiguation'. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, pp. 34–41.

Korhonen, Anna (2002). *Subcategorization acquisition*. Tech. rep. University of Cambridge Computer Laboratory.

Korhonen, Anna and Ted Briscoe (2004). 'Extended lexical-semantic classification of English verbs'. In: *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*. Boston, MA, pp. 38–45.

Korhonen, Anna, Yuval Krymolowski and Ted Briscoe (2006). 'A large subcategorization lexicon for natural language processing applications'. In: *Proceedings of the 5th international conference on Language Resources and Evaluation*. Genoa, Italy.

Korhonen, Anna, Yuval Krymolowski and Nigel Collier (2008). 'The choice of features for classification of verbs in biomedical texts'. In: *Proceedings of the 22nd International Conference on Computational Linguistics*. Vol. 1. Manchester, UK, pp. 449–456.

Korhonen, Anna, Yuval Krymolowski and Zvika Marx (2003). 'Clustering polysemic subcategorization frame distributions semantically'. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 64–71.

Krifka, Manfred (1987). *Nominal reference and temporal constitution: Towards a semantics of quantity*. Tech. rep. Tübingen: Universität Tübingen.

— (1992). 'Thematic relations as links between nominal reference and temporal constitution'. In: *Lexical matters*. Ed. by Ivan A. Sag and Anna Szabolcsi. Stanford, CA: Center for the Study of Language (CSLI). Chap. 2, pp. 29–53.

— (1998). 'The origins of telicity'. In: *Events and grammar*. Ed. by Susan Rothstein. Dordrecht: Kluwer, pp. 197–235.

Krovetz, Robert (1997). 'Homonymy and polysemy in information retrieval'. In: *Proceedings of the Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, pp. 72–79.

Kübler, Sandra (2008). 'The PaGe 2008 shared task on parsing German'. In: *Proceedings of the Workshop on Parsing German*. Columbus, OH: Association for Computational Linguistics, pp. 55–63.

Kučera, Henry and W. Nelson Francis (1967). 'Computational analysis of present-day American English'. Providence, RI: Brown University Press.

Kullback, Solomon and Richard A. Leibler (1951). 'On information and sufficiency'. In: *Annals of Mathematical Statistics* 22.1, pp. 79–86.

Lakoff, George (1972). 'Linguistics and natural logic'. In: *Semantics of Natural Language*. Ed. by Donald Davidson and Gilbert Harman. Dordrecht: Springer Netherlands, pp. 545–665.

Landauer, Thomas K. and Susan T. Dumais (1997). 'A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge'. In: *Psychological Review* 104.2, pp. 211–240.

Landis, J. Richard and Gary G. Koch (1977). 'The measurement of observer agreement for categorical data'. In: *Biometrics* 33.1, pp. 159–174.

Lang, Joel and Mirella Lapata (2010). 'Unsupervised induction of semantic roles'. In: *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, CA: Association for Computational Linguistics, pp. 939–947.

Langacker, Ronald W. (1987). 'Nouns and verbs'. In: *Language* 63.1, pp. 53–94.

Lee, Lillian (1999). 'Measures of distributional similarity'. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. College Park, MD: Association for Computational Linguistics, pp. 25–32.

Lenci, Alessandro (2010). 'Carving verb classes from corpora'. In: *Word Classes*. Ed. by Raffaele Simone and Francesca Masini. Amsterdam: John Benjamins.

Lenci, Alessandro, Gabriella Lapesa and Giulia Bonansinga (2012). 'LexIt: A computational resource on Italian argument structure'. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3712–3718.

Lenci, Alessandro, Barbara McGillivray, Simonetta Montemagni and Vito Pirrelli (2008). 'Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora'. In: *Proceedings of the Language Resources and Evaluation Conference*.

Levin, Beth (1993). 'English verb classes and alternations: A preliminary investigation'. Chicago: University of Chicago Press.

— (2011). 'Lexical conceptual structure'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Vol. 1. Berlin: Mouton de Gruyter. Chap. 19, pp. 420–440.

Levin, Beth and Malka Rappaport Hovav (1995). 'Unaccusativity: At the syntax-lexical semantics interface'. Cambridge, MA: MIT Press.

— (2005). 'Argument realization'. Cambridge: Cambridge University Press.

Li, Hang and Naoki Abe (1998). 'Generalizing case frames using a thesaurus and the MDL principle'. In: *Computational Linguistics* 24.2, pp. 217–244.

Li, Jianguo and Chris Brew (2008). 'Which are the best features for automatic verb classification'. In: *Proceedings of ACL-08: HLT*. Columbus, OH: Association for Computational Linguistics, pp. 434–442.

Light, Marc and Warren Greiff (2002). 'Statistical models for the induction and use of selectional preferences'. In: *Cognitive Science* 26.3, pp. 269–281.

Lin, Jianhua (1991). 'Divergence measures based on the Shannon entropy'. In: *IEEE Transactions on Information Theory* 37.1, pp. 145–151.

Link, Godehard (1983). 'The logical analysis of plurals and mass terms: A lattice-theoretic approach'. In: *Meaning, use and interpretation of language*. Ed. by Rainer Bäuerle, Christoph Schwarze and Arnim von Stechow. Berlin, Boston: De Gruyter, pp. 303–329.

Liu, Mei-chun and Ting-yi Chiang (2008). 'The construction of Mandarin VerbNet: A frame-based study of statement verbs'. In: *Language and Linguistics* 9.2, pp. 239–270.

Lloyd, Stuart (1982). 'Least squares quantization in PCM'. In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137.

Löbner, Sebastian (2013). 'Understanding semantics'. 2nd ed. London and New York, NY: Routledge.

Looks, Moshe, Marcello Herreshoff, DeLesley Hutchins and Peter Norvig (2017). 'Deep learning with dynamic computation graphs'. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR-17)*. arXiv: 1702.02181.

Loper, Edward, Szu-Ting Yi and Martha Palmer (2007). 'Combining lexical resources: Mapping between PropBank and VerbNet'. In: *Proceedings of the 7th International Workshop on Computational Linguistics*. Tilburg, the Netherlands.

Losonsky, Michael, ed. (1999). 'Humboldt: 'On language': On the diversity of human language construction and its influence on the mental development of the human species'. Oxford: Oxford University Press.

Louis, Annie and Ani Nenkova (2011). 'Automatic identification of general and specific sentences by leveraging discourse annotations'. In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 605–613.

Luhn, Hans Peter (1957). 'A statistical approach to mechanized encoding and searching of literary information'. In: *IBM Journal of Research and Development* 1.4, pp. 309–317.

Maienborn, Claudia (2011). 'Event semantics'. In: *Semantics: An international handbook of natural language meaning*. Ed. by Claudia Maienborn, Klaus von Heusinger and Paul Portner. Vol. 1. Berlin: Mouton de Gruyter. Chap. 34, pp. 802–829.

Manning, Christopher D. (1993). 'Automatic acquisition of a large subcategorization dictionary from corpora'. In: *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*. Columbus, OH: Association for Computational Linguistics, pp. 235–242.

Manning, Christopher D. and Hinrich Schütze (1999). 'Foundations of statistical natural language processing'. Cambridge, MA: MIT Press.

Maragoudakis, Manolis, Katia Lida Kermanidis and George Kokkinakis (2000). 'Learning subcategorization frames from corpora: A case study for modern Greek'. In: *Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*, pp. 19–22.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz (1993). 'Building a large annotated corpus of English: The Penn Treebank'. In: *Computational Linguistics* 19.2, pp. 313–330.

Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski and Suzanne Stevenson (2008). 'Semantic role labeling: An introduction to the special issue'. In: *Computational Linguistics* 34.2, pp. 145–159.

Mathew, Thomas A. and E. Graham Katz (2009). 'Supervised categorization for habitual versus episodic sentences'. In: *The Sixth Midwest Computational Linguistics Colloquium*.

McCallum, Andrew (2002). 'MALLET: A machine learning for language toolkit'.

McCarthy, Diana (2001). 'Lexical acquisition at the syntax-semantics interface: Diathesis alternations, subcategorization frames and selectional preferences'. PhD thesis. University of Sussex.

McCarthy, Diana and John Carroll (2003). 'Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences'. In: *Computational Linguistics* 29.4, pp. 639–654.

McCawley, James D. (1968). 'Lexical insertion in a transformational grammar without deep structure'. In: *Proceedings from the Annual Meeting of the Chicago Linguistic Society*. Vol. 4. 1. Chicago Linguistic Society, pp. 71–80.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov and Jan Hajič (2005). 'Non-projective dependency parsing using spanning tree algorithms'. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 523–530.

McNemar, Quinn (1947). 'Note on the sampling error of the difference between correlated proportions or percentages'. In: *Psychometrika* 12.2, pp. 153–157.

McRae, Ken, Michael J. Spivey-Knowlton and Michael K. Tanenhaus (1998). 'Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension'. In: *Journal of Memory and Language* 38.3, pp. 283–312.

Meilă, Marina and Jianbo Shi (2001). 'A random walks view of spectral segmentation'. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Merlo, Paola and Suzanne Stevenson (2001). 'Automatic verb classification based on statistical distributions of argument structure'. In: *Computational Linguistics* 27.3, pp. 373–408.

Merlo, Paola and Lonneke Van Der Plas (2009). 'Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both?' In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, pp. 288–296.

Messiant, Cédric and Thierry Poibeau (2008). 'LexSchem: A large subcategorization lexicon for French verbs'. In: *Proceedings of the Language Resources and Evaluation Conference*. Marrakech.

Metheniti, Eleni, Tim Van de Cruys and Nabil Hathout (2020). 'How relevant are selectional preferences for transformer-based language models?' In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1266–1278.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean (2013). 'Efficient estimation of word representations in vector space'. In: *Proceedings of Workshop at ICLR*.

Moens, Marc and Mark Steedman (1988). 'Temporal ontology and temporal reference'. In: *Computational Linguistics* 14.2, pp. 15–28.

Montemagni, Simonetta, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani and Remo Raffaelli (2003). 'Building the Italian syntactic-semantic treebank'. In: *Treebanks*. Ed. by Anne Abeillé. Dordrecht: Springer, pp. 189–210.

Mousser, Jaouad (2010). 'A large coverage verb taxonomy for Arabic'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

— (2011). 'Classifying Arabic verbs using sibling classes'. In: *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Mújdricza-Maydt, Éva, Silvana Hartmann, Iryna Gurevych and Anette Frank (2016). 'Combining semantic annotation of word sense & semantic roles: A novel annotation scheme for VerbNet roles on German language data'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3031–3038.

Nedjalkov, Vladimir P. and Sergej Je. Jaxontov (1988). 'The typology of resultative constructions'. In: *Typology of resultative constructions*. Ed. by Vladimir P. Nedjalkov and Bernard Comrie. Amsterdam: John Benjamins. Chap. 1, pp. 3–62.

Ng, Andrew Y., Michael I. Jordan and Yair Weiss (2002). 'On spectral clustering: Analysis and an algorithm'. In: *Advances in Neural Information Processing Systems*, pp. 849–856.

Ó Séaghdha, Diarmuid (2010). 'Latent variable models of selectional preference'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 435–444.

Ó Séaghdha, Diarmuid and Ann Copestake (2008). 'Semantic classification with distributional kernels'. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, pp. 649–656.

Ó Séaghdha, Diarmuid and Anna Korhonen (2012). 'Modelling selectional preferences in a lexical hierarchy'. In: *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*. Montréal, Canada, pp. 170–179.

Padó, Sebastian, Ulrike Padó and Katrin Erk (2007). 'Flexible, corpus-based modelling of human plausibility judgements'. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic, pp. 400–409.

Pala, Karel and Aleš Horák (2008). 'Can complex valency frames be universal?' In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*, pp. 41–48.

Palmer, Alexis, Elias Ponvert, Jason Baldridge and Carlota S. Smith (2007). 'A sequencing model for situation entity classification'. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 896–903.

Palmer, Frank R. (2001). 'Mood and modality'. 2nd ed. Cambridge: Cambridge University Press.

Palmer, Martha, Daniel Gildea and Paul Kingsbury (2005). 'The proposition bank: An annotated corpus of semantic roles'. In: *Computational Linguistics* 31.1, pp. 71–106.

Parisien, Christopher and Suzanne Stevenson (2010). 'Learning verb alternations in a usage-based Bayesian model'. In: *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci 2010)*. Portland, OR: Cognitive Science Society, pp. 2674–2680.

— (2011). 'Generalizing between form and meaning using learned verb classes'. In: *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci 2011)*. Boston, MA: Cognitive Science Society, pp. 2024–2030.

Paulus, Romain, Caiming Xiong and Richard Socher (2017). 'A deep reinforced model for abstractive summarization'. In: *arXiv preprint arXiv:1705.04304*. arXiv: 1705.04304.

Pennington, Jeffrey, 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)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

Perlmutter, David M. (1978). 'Impersonal passives and the unaccusative hypothesis'. In: *Annual Meeting of the Berkeley Linguistics Society*, pp. 157–190.

Peterson, Daniel W., Jordan Boyd-Graber, Martha Palmer and Daisuke Kawahara (2016). 'Leveraging VerbNet to build corpus-specific verb clusters'. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Berlin, Germany: Association for Computational Linguistics, pp. 102–107.

Petrov, Slav, Leon Barrett, Romain Thibaux and Dan Klein (2006). 'Learning accurate, compact, and interpretable tree annotation'. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, pp. 433–440.

Petrov, Slav and Dan Klein (2007). 'Improved inference for unlexicalized parsing'. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computa-*

tional Linguistics; Proceedings of the Main Conference. Rochester, NY: Association for Computational Linguistics, pp. 404–411.

Pianta, Emanuele, Luisa Bentivogli and Christian Girardi (2002). 'MultiWordNet: Developing an aligned multilingual database'. In: *Proceedings of the First International Conference on Global WordNet*, pp. 21–25.

Polajnar, Tamara and Stephen Clark (2014). 'Improving distributional semantic vectors through context selection and normalisation'. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 230–238.

Pollard, Carl and Ivan A. Sag (1994). 'Head-driven phrase structure grammar'. Chicago, IL: University of Chicago Press.

Preiss, Judita, Ted Briscoe and Anna Korhonen (2007). 'A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora'. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, pp. 912–919.

Pustejovsky, James, Kiyong Lee, Harry Bunt and Laurent Romary (2010). 'ISO-TimeML: An international standard for semantic annotation'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA).

Pustejovsky, James et al. (2003). 'The TIMEBANK corpus'. In: *Proceedings of Corpus Linguistics 2003*, pp. 647–656.

Puzicha, Jan, Thomas Hofmann and Joachim M. Buhmann (2000). 'A theory of proximity based clustering: Structure detection by optimization'. In: *Pattern Recognition* 33.4, pp. 617–634.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech and Jan Svartvik (1985). 'A comprehensive grammar of the English language'. London: Longman.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever (2019). 'Language models are unsupervised multitask learners'.

Radford, Andrew (1997). 'Syntax: A minimalist introduction'. Cambridge: Cambridge University Press.

Rafferty, Anna N. and Christopher D. Manning (2008). 'Parsing three German treebanks: Lexicalized and unlexicalized baselines'. In: *Proceedings of the Workshop on Parsing German*. Columbus, OH: Association for Computational Linguistics, pp. 40–46.

Rambelli, Giulia, Gianluca Lebani, Laurent Prévot and Alessandro Lenci (2016). 'LexFr: Adapting the LexIt framework to build a corpus-based French subcategorization lexicon'. In: *Proceedings of the Tenth International Conference on Language Resources and Evalu-*

ation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), pp. 930–937.

Raza, Ghulam (2011). 'Subcategorization acquisition and classes of predication in Urdu'. PhD thesis. Universität Konstanz.

Rehbein, Ines, Josef Ruppenhofer, Caroline Sporleder and Manfred Pinkal (2012). 'Adding nominal spice to SALSA: Frame-semantic annotation of German nouns and verbs'. In: *Proceedings of the 11th Conference on Natural Language Processing (KONVENS'12)*, pp. 89–97.

Reichart, Roi and Anna Korhonen (2013). 'Improved lexical acquisition through DPP-based verb clustering'. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 862–872.

Reichenbach, Hans (1947). 'Elements of symbolic logic'. Berkeley: University of California Press.

Resnik, Philip (1993). 'Selection and information: A class-based approach to lexical relationships'. PhD thesis. University of Pennsylvania.

— (1997). 'Selectional preference and sense disambiguation'. In: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*. Washington, DC, pp. 52–57.

Rissanen, Jorma (1978). 'Modeling by shortest data description'. In: *Automatica* 14, pp. 465–471.

Ritter, Alan, Mausam and Oren Etzioni (2010). 'A latent Dirichlet allocation method for selectional preferences'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 424–434.

Roberts, Will (2011). 'Integrating syntax and semantics for word sense disambiguation'. MA thesis. Saarbrücken, Germany: Universität des Saarlandes.

Roberts, Will and Markus Egg (2014). 'A comparison of selectional preference models for automatic verb classification'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 511–522.

Roberts, Will, Markus Egg and Valia Kordoni (2014). 'Subcategorisation acquisition from raw text for a free word-order language'. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 298–307.

Rohde, Douglas L. T. (2001). 'Tgrep2 User Manual'.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll and Franz Beil (1999). 'Inducing a semantically annotated lexicon via EM-based clustering'. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguist-*

ics. College Park, MD: Association for Computational Linguistics, pp. 104–111.

Rosenberg, Andrew and Julia Hirschberg (2007). 'V-measure: A conditional entropy-based external cluster evaluation measure'. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 410–420.

Roth, Michael and Mirella Lapata (2015). 'Context-aware frame-semantic role labeling'. In: *Transactions of the Association for Computational Linguistics* 3, pp. 449–460.

Ryle, Gilbert (1949). 'The concept of mind'. London: Barnes & Noble.

Saeed, John I. (1997). 'Semantics'. Oxford: Blackwell.

Sahlgren, Magnus (2006). 'The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces'. PhD thesis. Stockholm University.

Salton, Gerard and Michael J. McGill (1983). 'Introduction to Modern Information Retrieval'. New York, NY: McGraw-Hill.

Sarkar, Anoop and Daniel Zeman (2000). 'Automatic extraction of subcategorization frames for Czech'. In: *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, pp. 691–697.

Scarton, Carolina, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer and Anna Korhonen (2014). 'Verb clustering for Brazilian Portuguese'. In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 25–39.

Schabes, Yves, Anne Abeillé and Aravind K. Joshi (1988). 'Parsing strategies with 'lexicalized' grammars: Application to tree adjoining grammars'. In: *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88)*. Association for Computational Linguistics, pp. 578–583.

Scheible, Silke, Sabine Schulte im Walde, Marion Weller and Max Kisselew (2013). 'A compact but linguistically detailed database for German verb subcategorisation relying on dependency parses from Web corpora: Tool, guidelines and resource'. In: *Proceedings of the 8th Web as Corpus Workshop*. Lancaster, UK.

Schiller, Anne, Simone Teufel and Christine Thielen (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Tech. rep. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, and Seminar für Sprachwissenschaft, Universität Tübingen.

Schmid, Helmut (1994). 'Probabilistic part-of-speech tagging using decision trees'. In: *Proceedings of International Conference on New Methods in Language Processing*. Vol. 12, pp. 44–49.

— (1995). 'Improvements in part-of-speech tagging with an application to German'. In: *Proceedings of the EACL-SIGDAT Workshop*. Dublin, Ireland.

Schmid, Helmut, Arne Fitschen and Ulrich Heid (2004). 'SMOR: A German computational morphology covering derivation, composition and inflection'. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA).

Schulte im Walde, Sabine (2000). 'Clustering verbs semantically according to their alternation behaviour'. In: *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 747–753.

— (2002a). 'A subcategorisation lexicon for German verbs induced from a lexicalised PCFG'. In: *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*. Las Palmas, Canary Islands, Spain, pp. 1351–1357.

— (2002b). 'Evaluating verb subcategorisation frames learned by a German statistical grammar against manual definitions in the Duden Dictionary'. In: *Proceedings of the 10th EURALEX International Congress*, pp. 187–197.

— (2003). 'Experiments on the automatic induction of German semantic verb classes'. PhD thesis. Universität Stuttgart.

— (2006). 'Experiments on the automatic induction of German semantic verb classes'. In: *Computational Linguistics* 32.2, pp. 159–194.

— (2009). 'The induction of verb frames and verb classes from corpora'. In: *Corpus linguistics: An international handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 2. Berlin: Mouton de Gruyter. Chap. 44, pp. 952–971.

Schulte im Walde, Sabine and Chris Brew (2002). 'Inducing German semantic verb classes from purely syntactic subcategorisation information'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, pp. 223–230.

Schulte im Walde, Sabine, Christian Hying, Christian Scheible and Helmut Schmid (2008). 'Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences'. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, pp. 496–504.

Schumacher, Helmut (1986). 'Verben in Feldern'. Berlin, Boston: De Gruyter.

Schütze, Hinrich (1993). 'Word space'. In: *Advances in Neural Information Processing Systems*. Ed. by S Hanson, J Cowan and C Giles. Vol. 5. Morgan-Kaufmann, pp. 895–902.

Seddah, Djamé et al. (2013). 'Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically

rich languages'. In: *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. Seattle, WA: Association for Computational Linguistics, pp. 146–182.

Seeker, Wolfgang, Bernd Bohnet, Lilja Øvrelid and Jonas Kuhn (2010). 'Informed ways of improving data-driven dependency parsing for German'. In: *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, pp. 1122–1130.

Sennrich, Rico, Barry Haddow and Alexandra Birch (2016). 'Edinburgh neural machine translation systems for WMT 16'. In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, pp. 371–376.

Shi, Lei and Rada Mihalcea (2005). 'Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing'. In: *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico: Springer, pp. 100–111.

Shutova, Ekaterina, Lin Sun and Anna Korhonen (2010). 'Metaphor identification using verb and noun clustering'. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China: Association for Computational Linguistics, pp. 1002–1010.

Shutova, Ekaterina, Simone Teufel and Anna Korhonen (2013). 'Statistical metaphor processing'. In: *Computational Linguistics* 39.2, pp. 301–353.

Siegel, Eric V. and Kathleen R. McKeown (2000). 'Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights'. In: *Computational Linguistics* 26.4, pp. 595–627.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit (1997). 'An annotation scheme for free word order languages'. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, DC, pp. 88–95.

Smith, Carlota S. (1991). 'The parameter of aspect'. Dordrecht: Kluwer.

Smith, George (2003). *A brief introduction to the TIGER treebank, version 1*. Tech. rep. Universität Potsdam.

Spärck Jones, Karen (1972). 'A statistical interpretation of term specificity and its application in retrieval'. In: *Journal of Documentation* 28, pp. 11–21.

Spranger, Kristina and Ulrich Heid (2002). 'A Dutch chunker as a basis for the extraction of linguistic knowledge'. In: *Computational Linguistics in the Netherlands 2002, Selected Papers from the Thirteenth CLIN Meeting*. Ed. by Tanja Gaustad. Groningen: Rodopi, pp. 93–109.

Stevenson, Mark and Yorick Wilks (2001). 'The interaction of knowledge sources in word sense disambiguation'. In: *Computational Linguistics* 27.3, pp. 321–349.

Straka, Milan, Jan Hajič, Jana Straková and Jan Hajič, Jr (2015). 'Parsing universal dependency treebanks using neural networks and search-based oracle'. In: *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*. Ed. by Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk and Adam Przepiórkowski. Warsaw, Poland, pp. 208–220.

Strehl, Alexander, Joydeep Ghosh and Raymond Mooney (2000). 'Impact of similarity measures on web-page clustering'. In: *Proceedings of the Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pp. 58–64.

Subba, Rajen and Barbara Di Eugenio (2009). 'An effective discourse parser that uses rich linguistic information'. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, CO: Association for Computational Linguistics, pp. 566–574.

Sun, Lin (2012). 'Automatic induction of verb classes using clustering'. PhD thesis. Cambridge: University of Cambridge.

Sun, Lin and Anna Korhonen (2009). 'Improving verb clustering with automatically acquired selectional preferences'. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Singapore, pp. 638–647.

Sun, Lin, Anna Korhonen and Yuval Krymolowski (2008). 'Verb class discovery from rich syntactic data'. In: *Proceedings of the Ninth International Conference on Intelligent Text Processing and Computational Linguistics*. Haifa, Israel, pp. 16–27.

Sun, Lin, Anna Korhonen, Thierry Poibeau and Cédric Messiant (2010). 'Investigating the cross-linguistic potential of VerbNet-style classification'. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China: Association for Computational Linguistics, pp. 1056–1064.

Sun, Lin, Diana McCarthy and Anna Korhonen (2013). 'Diathesis alternation approximation for verb clustering'. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 736–741.

Surdeanu, Mihai, Sanda Harabagiu, John Williams and Paul Aarseth (2003). 'Using predicate-argument structures for information extraction'. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 8–15.

Swier, Robert S and Suzanne Stevenson (2004). 'Unsupervised semantic role labelling'. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, pp. 95–102.

Tai, Kai Sheng, Richard Socher and Christopher D. Manning (2015). 'Improved semantic representations from tree-structured long short-term memory 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)*. Beijing, China: Association for Computational Linguistics, pp. 1556–1566.

Telljohann, Heike, Erhard Hinrichs and Sandra Kübler (2004). 'The TüBa-D/Z treebank: Annotating German with a context-free backbone'. In: *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal: European Language Resources Association (ELRA), pp. 2229–2232.

Tenny, Carol (1987). 'Grammaticalizing aspect and affectedness'. PhD thesis. Massachusetts Institute of Technology.

Tesnière, Lucien (1959). 'Éleménts de syntaxe structurale'. Paris, France: Klincksieck.

Thater, Stefan, Hagen Fürstenau and Manfred Pinkal (2011). 'Word meaning in context: A simple and effective vector model'. In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 1134–1143.

Tishby, Naftali, Fernando C.N. Pereira and William Bialek (2000). 'The information bottleneck method'. In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*. Ed. by Bruce Hajek and R.S. Sreenivas, pp. 368–377.

Titov, Ivan and Alexandre Klementiev (2012). 'A Bayesian approach to unsupervised semantic role induction'. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 12–22.

Trueswell, John C., Michael K. Tanenhaus and Susan M. Garnsey (1994). 'Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution'. In: *Journal of Memory and Language* 33.3, pp. 285–318.

Turney, Peter D. and Patrick Pantel (2010). 'From frequency to meaning: Vector space models of semantics'. In: *Journal of Artificial Intelligence Research* 37.1, pp. 141–188.

Van de Cruys, Tim (2009). 'A non-negative tensor factorization model for selectional preference induction'. In: *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Athens, Greece: Association for Computational Linguistics, pp. 83–90.

— (2014). 'A neural network approach to selectional preference acquisition'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 26–35.

Vázquez, Gloria, Ana Fernández, Irene Castellón and Marìa Antonia Martì (2000). *Clasificacíon verbal: Alternancias de díatesis*. Tech. rep. Universitat de Lleida.

Vendler, Zeno (1957). 'Verbs and times'. In: *The Philosophical Review* 66, pp. 143–160.

— (1967). 'Linguistics in philosophy'. Ithaca, NY: Cornell University Press.

Verhagen, Marc, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz and James Pustejovsky (2007). 'SemEval-2007 task 15: TempEval temporal relation identification'. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 75–80.

Verkuyl, Henk J. (1972). 'On the compositional nature of the aspects'. Dordrecht: Reidel.

Viterbi, Andrew (1967). 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm'. In: *IEEE Transactions on Information Theory* 13.2, pp. 260–269.

Vlachos, Andreas, Anna Korhonen and Zoubin Ghahramani (2009). 'Unsupervised and constrained Dirichlet process mixture models for verb clustering'. In: *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Athens, Greece: Association for Computational Linguistics, pp. 74–82.

Vulić, Ivan, Nikola Mrkšić and Anna Korhonen (2017). 'Cross-lingual induction and transfer of verb classes based on word vector space specialisation'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2546–2558.

W3Techs.com (2020). 'Usage statistics of content languages for websites'.

Ward, Jr, Joe H. (1963). 'Hierarchical grouping to optimize an objective function'. In: *Journal of the American Statistical Association* 58.301, pp. 236–244.

Wauschkuhn, Oliver (1999). 'Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora'. Shaker Verlag.

Weller, Marion, Alexander Fraser and Sabine Schulte im Walde (2013). 'Using subcategorization knowledge to improve case prediction for translation to German'. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 593–603.

Wilks, Yorick (1975). 'An intelligent analyzer and understander of English'. In: *Communications of the ACM* 18.5, pp. 264–274.

— (1978). 'Making preferences more active'. In: *Artificial Intelligence* 11.3, pp. 197–223.

Wright, Georg Henrik von (1963). 'Norm and action'. New York, NY: Humanities Press.

Yarowsky, David (1993). 'One sense per collocation'. In: *Proceedings of the Workshop on Human Language Technology*, pp. 266–271.

Ye, Patrick and Timothy Baldwin (2006). 'Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler'. In: *Proceedings of the Australasian Language Technology Workshop*. Sydney, Australia, pp. 139–148.

Yeh, Alexander (2000). 'More accurate tests for the statistical significance of result differences'. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken, Germany, pp. 947–953.

Yu, Liang-Chih, Jin Wang, K. Robert Lai and Xuejie Zhang (2017). 'Refining word embeddings for sentiment analysis'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 534–539.

Zapirain, Beñat, Eneko Agirre and Lluís Màrquez (2008). 'Robustness and generalization of role sets: PropBank vs. VerbNet'. In: *Proceedings of ACL-08: HLT*. Columbus, OH: Association for Computational Linguistics, pp. 550–558.

Zapirain, Beñat, Eneko Agirre, Lluís Màrquez and Mihai Surdeanu (2013). 'Selectional preferences for semantic role classification'. In: *Computational Linguistics* 39.3, pp. 631–663.

Zarcone, Alessandra and Alessandro Lenci (2008). 'Computational models for event type classification in context'. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 1232–1238.

Zeller, Tom (2018). 'Detecting ambiguity in statutory texts'. Bachelor's Thesis. Universität Stuttgart.

Zhang, Zhuosheng, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou and Xiang Zhou (2020). 'Semantics-aware BERT for language understanding'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 5. New York, NY: AAAI Press, pp. 9628–9635.

Zipf, George Kingsley (1949). 'Human behavior and the principle of least effort: An introduction to human ecology'. Cambridge, MA: Addison-Wesley.

SELBSTSTÄNDIGKEITSERKLÄRUNG ZUR DISSERTATION

Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten Dissertation mit dem Titel

Facets of verb meaning:
A distributional investigation of German verbs

um eine von mir erstmalig, selbstständig und ohne fremde Hilfe verfasste Arbeit handelt.

Ich erkläre ausdrücklich, dass ich *sämtliche* in der oben genannten Arbeit verwendeten fremden Quellen, auch aus dem Internet (einschließlich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken u. Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken u. Ä. anderer Autorinnen und Autoren (Paraphrasen) die Quelle angegeben habe.

Mir ist bewusst, dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und nach §16 der Promotionsordnung der Sprach- und literaturwissenschaftlichen Fakultät entsprechend geahndet werden.

Berlin, 1 April 2022

William Roberts

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Final Version as of 21st April 2023 (`classicthesis` version 1.0).