Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German

Thomas M. Segler

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh
2007

Abstract

This thesis considers possible criteria for the selection of example sentences for difficult or unknown words in reading texts for students of German as a Second Language (GSL). The examples are intended to be provided within the context of an Intelligent Computer-Aided Language Learning (ICALL) Vocabulary Learning System, where students can choose among several explanation options for difficult words. Some of these options (e.g. glosses) have received a good deal of attention in the ICALL/Second Language (L2) Acquisition literature; in contrast, literature on examples has been the near exclusive province of lexicographers.

The selection of examples is explored from an educational, L2 teaching point of view: the thesis is intended as a first exploration of the question of what makes an example helpful to the L2 student from the perspective of L2 teachers. An important motivation for this work is that selecting examples from a dictionary or randomly from a corpus has several drawbacks: first, the number of available dictionary examples is limited; second, the examples fail to take into account the context in which the word was encountered; and third, the rationale and precise principles behind the selection of dictionary examples is usually less than clear.

Central to this thesis is the hypothesis that a random selection of example sentences from a suitable corpus can be improved by a guided selection process that takes into account characteristics of helpful examples.

This is investigated by an empirical study conducted with teachers of L2 German. The teacher data show that four dimensions are significant criteria amenable to analysis: (a) reduced syntactic complexity, (b) sentence similarity, provision of (c) significant co-occurrences and (d) semantically related words.

Models based on these dimensions are developed using logistic regression analysis, and evaluated through two further empirical studies with teachers and students of L2 German.

The results of the teacher evaluation are encouraging: for the teacher evaluation, they indicate that, for one of the models, the top-ranked selections perform on the same level as dictionary examples. In addition, the model provides a ranking of potential examples that roughly corresponds to that of experienced teachers of L2 German. The student evaluation confirms and notably improves on the teacher evaluation in that the best-performing model of the teacher evaluation significantly outperforms both random corpus selections and dictionary examples (when a penalty for missing entries is included).

Acknowledgements

First and foremost, I would like to thank my supervisors Helen Pain, Frank Keller, and Antonella Sorace. Without their consistent advice, support and guidance throughout the years this thesis would not have come about. Furthermore, I am highly indebted to Peter Wiemer-Hastings for providing me with invaluable advice, support and extremely useful comments. Thanks are also due to the many teachers and students who participated in the studies presented in this thesis.

I am also very grateful to my friends and colleagues both at South Bridge and Buccleuch Place, who helped make my time here in Edinburgh such an enjoyable experience. In particular, I would like to thank Ben Curry, Jacques Fleuriot and Ruli Manurung for their general support, encouragement and very helpful last minute proofreading. Finally, I would like to thank my parents for supporting me in every way throughout the completion of this thesis.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Thomas M. Segler*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

As anyone who has ever spent some time and effort on learning a second language will probably attest, dictionaries that contain example sentences tend to be considerably more helpful when encountering an unknown or difficult second language word than those that only provide a translation, or brief textual gloss (e.g. paraphrase) for that word. In many cases, the elucidation of the sense of the unknown word benefits considerably from an illustrative phrase or sentence. The popularity of example sentences (ES)[1] among second language learners is borne out in studies such as (Béjoint, 1981) who found that French learners of L2 English consulted examples and quotations more often than any other explanation option provided by monolingual general English dictionaries. The extensive use of ES is not surprising, considering that they can be used by the learner both for interpretation (decoding) and composition (encoding) (Cowie, 1980).

Considering the evident benefit that example sentences provide for second language learners, it is surprising that the criteria leading to their selection in dictionaries often seem to be shrouded in mystery. In the current thesis, the issue will therefore be addressed from a pedagogic, teacher-oriented angle: what makes an example sentence a helpful one for a second language learner, in the view of second language teachers?

Throughout this thesis, the second language serving as the object of investigation will be German. The focus of the current dissertation is on developing a computational model of the criteria that experienced teachers of German as a Second Language (L2 German)[2] employ for the selection of ES. The examples will be selected by the teachers for difficult or unknown target words in a given German reading text that they consider

[1] Throughout this thesis (except in sections 2.2 and 2.4), the terms 'example' and 'example sentence' will be used interchangeably.

[2] See section 1.1 on the use of this terminology throughout this thesis.

to be the most helpful in each case for their respective group of students. It should be emphasized at this point that the research presented in this thesis is intended as an exploratory step towards modeling the teacher criteria and an investigation of the issues involved.

As has been intimated above, ES in learner dictionaries essentially serve a dual function: that of *encoding*, i.e. exemplifying the *typical usage* of the target word, and that of *decoding*, i.e. that of clarifying the *meaning* of the word. It is the latter of these partial functions — that of decoding — that the current thesis focusses on. This narrowing of the focus was motivated by the following considerations: first and foremost, the task of identifying typical usages can be achieved relatively straightforwardly by lexicographers via analyses based on corpus linguistics, whereas native speaker judgments tend to be unreliable for this sort of task (Fox, 1987). Second, if a language learner comes across a target word whose meaning is unknown or at least unclear to him, the decoding function of ES is arguably of more pressing concern for him than knowing how he can productively use the word (in that sense, the decoding aspect of the ES's function may be considered a prerequisite of sorts to the encoding aspect). The decision to focus on the decoding function of ES means that, for the purposes of this research, the concept of an example being helpful for a language learner will be recast in more precise terms as an example being helpful *in illustrating the meaning of the target word* (as it was used in the original sentence (OS) of the reading text in question).

The model for example sentence selection to be developed in this thesis is intended to be provided within the context of an Intelligent Computer-Aided Language Learning (ICALL) system, where students can choose among several explanation options for difficult words. Apart from example sentences which are the object of investigation in this thesis, these are envisaged to include pictorial and textual glosses (e.g. definitions, translations, and paraphrases). Implicit in this categorization (as well as in the usual lexicographic definition of example sentences provided in section 2.2) is that example sentences are explanation options that are distinct from both definitions and paraphrases.

Unlike many existing lexical reading tutors, which — if they allow for the provision of example sentences at all — often only present either a single or a very limited preselected number of examples (e.g. Coady et al. (1993); Beheydt (1990)), the to-be-developed system is envisaged to be able to draw upon a vast number of available corpus examples as potential example selections.

In contrast to definitions, which may be regarded as *explicit* explanations of a target word, ES can explain the meaning only *implicitly* since they are understood to be produced in a "natural communication" setting, i.e. from one native speaker to another. It is worth noting at this point that the above holds true regardless of whether the example is actually found in a corpus of written or spoken language, i.e. is 'authentic' by virtue of being corpus-attested, or concocted by a lexicographer or teacher for didactic purposes. In both cases, the example sentence is produced with what Zöfgen (1994) has termed "quasi-communicative intention"[3].

In contrast to paraphrases of the target word, ES contain (at least one instance or token of) the word in question. In fact, *every* well-formed sentence that can be found or invented is a potential ES of a target word, the only constraint being the target word's occurrence in that sentence. It is this characteristic which renders ES selection a non-trivial task: even if one constrains the set of potential ES to authentic, corpus-attested sentences, a vast range of examples — depending on the frequency of the target word — are candidates for presentation to the language learner; if one also permits invented examples to be considered, the frequency constraint is removed and the range of potential ES becomes infinite. By way of contrast, textual glosses such as definitions and paraphrases or synonyms are relatively tightly constrained by their respective functions, and can be expected to differ from one another in relatively minor details only.

It is therefore self-evident that a random selection of example sentences from some corpus is far from the ideal solution for the task at hand, since it is very likely that other ES could be found that are more helpful to the language learner. Central to this thesis is the hypothesis that such a random selection of ES from a suitable corpus can be significantly improved by a guided selection process that takes into account characteristics of helpful examples. As the discussion in chapter 2 will show, there is no agreement in either the lexicographic discussion or the actual treatment in learner dictionaries regarding the issue of what exactly these characteristics should be, or for that matter on the related question of the ideal source and form of example sentences. Given this situation, the current thesis adopts a pedagogic, teacher-oriented approach to the issue, i.e. focusses on what *teachers* consider to be helpful examples.

It should be emphasized at this juncture that the thesis does not deal with the "context" or Word Sense Disambiguation (WSD) problem for example sentence selection that will have to be addressed by any future implementation of the system. As word

[3] "[...] in quasi-kommunikativer Absicht verfaßt [...]" (Zöfgen, 1994, p. 192).

meanings are for the most part at least to some extent defined by the context, the problem of excluding examples that contain the target word in an "inappropriate" sense (with respect to the use of the word in the original sentence context) is a non-trivial one (see section 6.2 for how the WSD problem is addressed in this thesis).

It should also be noted at this stage that the focus of the thesis is on analyzing the criteria that teachers employ in their example selection, and on developing a computational model based on this analysis. The wider-ranging issues of how the model can be implemented within the envisaged Vocabulary Learning Environment, whether it actually helps learners comprehending German reading texts or acquiring and retaining new vocabulary, or under what circumstances other explanation options such as definitions are more helpful to the learners, are outside the scope of this thesis. Instead, the to-be-developed models will be evaluated both in terms of their merits as teacher models, i.e. in terms of the degree to which they are able to model the teachers' selection criteria, and in terms of their perceived helpfulness to learners of L2 German.

The intended target group of the learners benefitting from the model are advanced-level students of L2 German, for whom syntax does not present an an insurmountable obstacle to reading and text comprehension; since the teachers participating in the exploratory study (chapter 3) were based in Scotland, the assumed L1 of the target group of learners is English.

1.1 Some Terminology

This section will briefly clarify some of the terminology to be used throughout this thesis.

- The terms **Second Language (Learning/Acquisition)** (or **L2**), and **Foreign Language (Learning/Acquisition)** tend not to be used in a consistent manner in the Applied Linguistics literature; the situation is made even more confusing by the fact that many learner dictionaries tend to refer to the learned language as a *foreign language*, whereas the general term of the research field is *Second Language Acquisition/Learning*. To the extent that a deliberate distinction between the terms is made, *second language* tends to imply that "the language plays an institutional and social role in the community", whereas the implication of *Foreign Language Learning* is that it "takes place in settings where the language plays no major role in the community and is primarily learnt only in

the classroom" (Ellis, 1994, p. 12). Sometimes, *second language* also carries the implication that the language is learnt as the *second* (as opposed to third etc.) foreign language.

Given the lack of a separate neutral and superordinate term to cover all above-mentioned types of learning, **Second Language (Learning)** (or **L2**) will be used for this purpose throughout this thesis (i.e. none of the aforementioned specific implications is intended by its use).

- In keeping with the intuitive understanding and definition of the term **word**, a word (or **target word** when referring to the unknown or difficult word in need of explanation to the learner) is taken to be "any segment of written or printed discourse ordinarily appearing between spaces or between a space and a punctuation mark" (*Merriam Webster Online*). Throughout the thesis, it will be clear from the context whether a particular usage of *word* refers to a word *form* or the *sense* of the word (the concept denoted by the word in the given context); i.e. saying that a target word encountered in a reading text is unknown to a language learner means that its corresponding sense (in the given context) is unknown to him.

- The term *Example Sentence* implies the usual intuitive meaning of the term 'sentence', i.e. a well-formed stand-alone grammatical unit of one or more clauses consisting of at least a subject and a verb, surrounded by punctuation marks. This means that, for the purposes of this thesis, the following do not count as valid ES:

  - *Sentence fragments* or phrases sometime provided by dictionaries in lieu of full example sentences (e.g. "the *buildings* in the old town" as an example for *building*);
  - Two or more consecutive sentences to explain a target word.

1.2 Setting the Scene: Motivation for the Thesis

An important motivation for the research presented in this thesis is that for the intended task at hand (i.e. presenting ES in the context of a Vocabulary Learning Environment where learners can click on difficult or unknown target words in reading texts), se-

lecting examples from a suitable lexicographic resource such as an electronic learner dictionary has several drawbacks.

First, the number of available dictionary ES is severely limited even if one includes examples from more than just one dictionary. This is especially the case if one considers that the coverage with respect to ES is incomplete even in the most comprehensive current learner dictionaries. However, it is clear that a language learner benefits from being exposed to as many examples for a word being used as possible: in contrast to definitions, only a few aspects of the word meaning can usually be inferred from the context provided by any one example — this tends to be true even for "forcing examples" where the context constrains possible meanings of the target word to such an extent that only one plausible interpretation as to the target word's meaning remains. So, as pointed out by Black (1991), "a series of examples may be necessary to fully illustrate a particular word"; in the same vein, Schouten-van Parreren (1989) and Beheydt (1987) argue for a presentation of new words in a variety of different, meaningful contexts to enable successful "semantisation" of new vocabulary.

Second, as has been noted at the beginning of this section, the selection criteria for dictionary ES are often far from transparent. Learner dictionaries tend to differ to a considerable extent on the issue of whether to use authentic (corpus-*attested*) examples, examples that are only corpus-*oriented* (i.e. possibly invented by the lexicographer, using corpus occurrences only as a guideline), or — in the extreme case — freely invented by the lexicographer as he sees fit (see also chapter 2 for a survey of the corresponding discussion in L2 lexicography).

Third, studies on dictionary use such as (Marello, 1987), (Black, 1991) and (Nesi, 1996) suggest the possibility that dictionary examples often fail to provide the sort of information that learners need.

Finally, the ES provided by dictionaries by necessity fail to take into consideration the context in which the word was encountered by the learner in the reading text: one may well hypothesize that a language learner reading an L2 text benefits from an example that is semantically similar to the original sentence containing the target word (i.e. is "about the same topic").

It has already been pointed out that a random selection from corpus examples is not a viable alternative; instead, the guiding motivation for this thesis is to investigate how a random selection can be improved by a model of criteria that experienced teachers of L2 German employ in their selection of ES. On the basis of the ranking the model assigns to potential example sentences, a learner could then be provided with a set of

examples that best reflect the selection criteria of L2 teachers.

1.3 Aims of this Research

The overall aim of the research project presented in this dissertation was to build a computational model which reflects criteria for the selection of ES for difficult or unknown target words in reading texts of L2 German. In order to achieve this aim, an empirical study was conducted where experienced teachers of L2 German were asked to provide invented examples they considered most helpful for their students. Based on the analysis of the teacher data, the goal was to develop a model of the teacher selection criteria based on regression analysis. Finally, the resulting models were to be evaluated through empirical studies with both teachers and learners of L2 German.

More specifically, the research had the following aims:

- To elicit ES in the form of invented examples from experienced teachers of L2 German for target words in a reading text that they believed to be most helpful for illustrating or clarifying the meaning of the word (for the given reading context and for their respective group of students);

- To gather explanations from the teachers as to the reasons why they considered their examples particularly helpful;

- To analyze whether the criteria suggested by the teachers' explanations were significant factors for their selection of examples;

- To determine which of the various measures suggested in the literature for the various analysis dimensions is the best choice for the respective analysis tasks;

- To develop logistic regression models based on the analysis dimensions found to be significant ES selection factors in the above-mentioned analyses;

- To evaluate the resulting models in two empirical studies with teachers and learners of L2 German, respectively; in particular, to address the questions of (a) how the models' top-ranked selections compare to the gold standard of examples provided by an experienced teacher of L2 German, to dictionary examples, and to

each other; and (b) whether or not the models provide consistent internal ordering, i.e. whether the models' top-ranked selections were rated as significantly more helpful than both random corpus selections and bottom-ranked examples.

The following sections of this thesis will address these specific aims and attempt to establish the extent to which they have been met.

1.4 Structure of the Thesis

The remainder of this thesis comprises eight chapters the content of which is summarized below.

- **Chapter 2** presents some relevant background of the L2 lexicographic literature on example sentences. It surveys the main issues of the lexicographic discussion on desirable sources, forms and functions of example sentences in learner dictionaries.

- **Chapter 3** presents the design of the exploratory study in which teachers of L2 German were asked to provide helpful examples and explanations thereof; the chapter also contains an analysis of the explanation section of the teacher data, which motivates the choice of potential ES selection factors analyzed in chapters 4 to 6.

- **Chapter 4** surveys and empirically evaluates several measures of syntactic complexity. The measures are evaluated on their correlations with native speaker judgments on syntactic complexity; the measure with the highest correlation — sentence length — is then used to show that the syntactic complexity of the teachers' ES had been significantly reduced compared to the corresponding OS. The chapter also motivates the use of syntactic complexity as a pre-filter for the model of teacher criteria to be developed in chapter 7.

- **Chapter 5** considers the criterion of sentence similarity of the OS/ES sentence pairs. It surveys several measures of sentence similarity suggested in the literature; of these, it selects and empirically evaluates two measures — lexical overlap and Latent Semantic Analysis (LSA) — with respect to their correlations with human judgments of sentence similarity. The chapter shows that LSA yields remarkably high correlations with human sentence similarity judgments,

motivating its use as the measure for sentence similarity among the OS/ES pairs of the teacher data. In more specific terms, the chapter shows that the degree of sentence similarity found in the teacher data is significantly higher than that found in randomly selected sentence pairs.

- **Chapter 6** is concerned with specific lexical choices teachers may have made in their ES selection. In particular, the chapter shows that both paradigmatic lexical relations and significant co-occurrences are significant factors in the teacher examples, in contrast to choices relating to the morphological form of the target word. In the context of the analysis of paradigmatic lexical relations, the chapter surveys several measures of word similarity, of which LC-IR as a statistical web-based approach and LSA are selected and compared on the tasks of a multiple-choice lexical relation test, and a correlation analysis with native speaker similarity ratings of German noun pairs. On this issue, the chapter concludes that LSA's and LC-IR's performances are inconclusive and less-than-optimal for the task at hand, motivating the use of a manual, dictionary-guided approach to measuring word similarity.

- **Chapter 7** presents three different logistic regression models of the teachers' ES selection criteria based on the above-mentioned factors that had proved significant in the selection of the teacher examples. The models developed in this chapter differ in whether or not they exclude difficult vocabulary.

- **Chapter 8** contains the evaluation of the models in the form of two studies in which both experienced teachers and intermediate-to-advanced students of L2 German were asked to rate the helpfulness of the models' preferred examples compared to each other, as well as vis-à-vis the gold standard of teacher-provided examples and dictionary examples. The question of whether the models provide consistent internal ordering, and whether the models' top-ranked examples are rated as significantly more helpful than both random corpus selections and the models' bottom-ranked examples, is also addressed in the evaluation.

- **Chapter 9** presents a summary of the issues explored in the current thesis; it provides the author's conclusions as to the limitations and contributions of the research presented; and it discusses the possible future work that can be pursued in relation to the work discussed in this thesis.

Chapter 2

Background: Example Sentences in L2 Lexicography

2.1 Introduction

It has been argued in the preceding chapter that selecting example sentences from (learner or general purpose) dictionaries has several significant shortcomings for the envisaged application of reading texts of L2 German in the framework of an ICALL Vocabulary Learning Environment. Some of these drawbacks are quite obvious, such as the limited number of available examples, and the failure to take into account the original reading context. This chapter is concerned with some background issues relating to the selection of examples in learner dictionaries that are in need of further elucidation. In particular, this chapter will survey the relevant metalexicographic discussion of desiderata for sources, forms and functions of dictionary examples, as well as the actual lexicographic practice of treating example sentences in current dictionaries of L2 German. This will help to motivate the pedagogic, teacher-centered approach adopted in the remainder of this thesis, and to set the scene for the empirical study on the criteria for the selection of example sentences employed by teachers of L2 German in chapter 3.

As most language learners will probably attest, existing dictionaries appear to be inconsistent in their exemplification of target words, both with respect to the type of exemplification (full sentence or short phrase), the number of provided examples per entry, and whether or not an entry contains any exemplification at all. The dictionaries' preface sections tend not to shed much light on the issue either: in most cases, either the selection criteria for examples are not mentioned at all, or some vague reference

to a corpus being used as a guideline is made. Given this situation, the remainder of this chapter will first attempt to shed some light on the concept of an example sentence as used in learner dictionaries, and clarify its status within a typical dictionary entry vis-à-vis a definition. The chapter then surveys the metalexicographic discussion on desirable features of examples. Finally, an overview is given on how the metalexicographic discussion is reflected in actual lexicographic practice as attested by current dictionaries of L2 German. The chapter concludes with a summary of the discussion, and its implications for the remainder of this thesis.

2.2 What are Lexicographic Examples?

Before we turn to the metalexicographic discussion of which *forms* of examples are desirable, what their *sources* should be, and which *functions* they should fulfill, mention should be made of what defines a dictionary example according to the metalexicographic literature. Hartmann (2001, p. 173) offers the following definition of a dictionary example: "A word or phrase used in a dictionary or other reference work to illustrate a form or meaning in a wider context, either as a citation excerpted from a text corpus or as a specimen invented by the compiler." Creamer (1987, p. 241) states that "the primary purpose of an example is to demonstrate the use of a word in its natural environment. An example can be either a few words, a sentence pattern, or a complete sentence."

Several aspects are noteworthy about these definitions: first, the illustrative function of examples can refer to either *form* or *meaning*, i.e. examples may be used to illustrate both grammatical patterning (e.g. indicate collocations) and sense. For the research presented in this thesis, and the discussion in the remainder of this chapter, it is mainly the latter, meaning-related aspect that will be of interest.

Second, emphasis is placed on the context-providing, *illustrative* aspect of examples (either with regard to form or meaning), which is achieved by situating the target word in a wider 'micro-context'. Creamer's definition also alludes to the aspect of *naturalness* often invoked as a criterion for lexicographic examples (see section 2.5).

Third, the question of how wide the surrounding context should be (i.e. the *length* of the example), is left unspecified: anything from a single word to longer verbal or non-verbal phrases (possibly full sentences) is seen as a valid lexicographic exemplification.[1]

[1]For the target word *omen*, a phrasal example would be *a bird of ill omen*, while *It is an omen of*

Fourth, two possible *sources* for examples are mentioned: they could be either citations from a text corpus (so-called corpus-attested or "authentic" examples), or created by the lexicographer as he or she sees fit ("invented" examples). The latter two points are taken up in the following section.

Finally, the above definitions, taken together, indicate that lexicographic examples fulfill a dual function: while Hartmann seems to emphasize the *receptive* (decoding) use of examples, Creamer's definition appears to place greater importance on the *productive* (encoding) needs of the dictionary user by referring to the possible *usages* of the target word (see also the discussion in section 2.6). As was noted in the preceding chapter, it is the former, decoding function of examples that the current thesis will focus on.

2.3 The Role of Examples in Dictionary Entries

Martin's (1989) strong endorsement of example sentences as the most rewarding and useful part of a dictionary[2] tends not to be reflected in metalexicographic literature. Insofar as the role and relative status of examples in relation to other parts of the dictionary entries are mentioned at all, examples tend to be seen as playing an ancillary role to definitions and translations. The primacy of definitions is noted by Black (1991) and Creamer (1987) who see examples as supplementing and possibly extending the definitions. Creamer (1987, p. 241) argues that an example can "take the burden off a definition by showing various ways the entry can be translated in context, indicate typical modifiers, and illustrate points of usage (e.g., if the entry collocates with a certain verb)."

However, if one restricts the focus of attention to learner dictionaries, one finds that examples tend to be accorded a more important role: according to Herbst (1989, p. 1382), examples in learner dictionaries are "intended to illustrate the meanings of words more clearly than is sometimes possible within the definition [...]". In a similar vein, Jackson (2002) draws attention to the crucial role that examples appear to play in learner dictionaries, observing that these contain particularly numerous instances of examples.

---

success exemplifies a full-sentence example (both examples are taken from Collins (1991)).

[2] "La tentation est grande de conclure avec les Acadèmiciens (*Préface* 1878) que 'les exemples sont la vraie richesse et la partie la plus utile du dictionnaire"' (Martin, 1989, p. 606).

2.4 Form, Length and Number of Lexicographic Examples

We saw in section 2.2 that lexicographic exemplifications come in two guises: either a verbal or non-verbal phrase that includes the target word (possibly consisting of one word only), or a grammatically complete sentence that includes the target word. As was mentioned in chapter 1, it is only the latter type of complete example sentences that will be of interest for the following sections of this chapter, and indeed throughout this thesis.

The above being said, it should be noted that most dictionaries (both learner and general-purpose) use both types, yet typically remain silent on the criteria underlying the use of each type. This could be seen as a reflection of the lack of consensus in the metalexicographic literature on the question of the preferable form or length of examples.

The question of how long examples should be is often discussed in the same breath as the closely related question on whether to use phrasal or sentence-level examples. As the overview below will demonstrate, there is no consensus in the metalexicographic literature on the length or form of lexicographic examples.

Some lexicographers feel that short, below sentence-level examples may be preferable not only for reasons of space (Jacobsen et al., 1991), but also because long, sentence-level examples could come at the expense of clarity (Antor, 1994).[3] In the same vein, Nikula (1986, p. 190) argues that examples need not be longer than is necessary to fulfill their prototype-function in the dictionary context (see below). Zgusta generally seems to prefer phrases to sentences on the grounds that sentences tend to contain too much specific information that for the learner may be difficult to generalize to other possible constructions containing the target word. Zgusta also links the question of example length to the source of the example: if the lexicographer invents his own examples, then he recommends that the examples be "very short, for example, only the verb — its object, the adjective with the substantive or vice versa" (Zgusta, 1971, p. 267).

On the other hand, Cowie champions the inclusion of full example sentences, arguing that "they can be used to illustrate grammatical patterning and to provide sufficient

[3] As a case in point, Antor (1994, p. 80) cites the following example for the target word *mettle* (taken from *LDOCE*): *The runner fell and twisted his ankle badly, but he showed his* **mettle** *by continuing in the race.*

context for meaning and stylistic level to be clearly established" (Cowie, 1989, p. 57). Despite their preference for short examples, Jacobsen et al. (1991, p. 2788) concede that "often more than a micro-context is needed if an example is meant to illustrate a feature at sentence-level or above". Zöfgen (1986) goes even further by suggesting that the extra information provided by sentences is exactly what learners need; in fact, he recommends that example sentences contain as much implicit information about the target word as possible in order to provide sufficient clues on grammatical, collocational and other elements to the language learner.

In the L2 Vocabulary Acquisition Literature, this type of rich, supportive, "pregnant" contexts has been championed by researchers such as Beheydt (1987, 1990) and Schouten-van Parreren (1989). On the other hand, Mondria and Wit-De Boer (1991) have found that rich contexts may draw attention away from the lexical level and, while improving guessability and facilitating reading comprehension, are not conducive to retention.

Mondria and Wit-De Boer have found that the following types of factors determine guessability: contextual factors (redundancy of context, occurrence of lexical relations such as synonyms and antonyms), word factors (e.g. part-of-speech and transparency of word structure), and learner/reader factors.

As for the question how many examples should be provided, there is widespread agreement on the benefit of providing as many examples as possible (see the discussion in section 1.2). Maingay and Rundell (1987, p. 131) argue for a "series of well-chosen examples" to build up "a complete picture of a word's salient features, because this to some extent replicates — however imperfectly — the process of repeated exposure by which native speakers achieve their competence". Striking a similar note, Beheydt (1987, p. 64) argues that new words should be "used again and again in a variety of contexts", as "it is only in various meaningful contexts that the full polysemous versatility of the word is revealed, as well as its syntactic and morphological potential."

Nikula (1986); Zöfgen (1986); Lenz (1998) argue that 'authentic' examples, i.e. citations, need to be provided in larger numbers than invented examples if they are to fulfill their L2-relevant functions, because they are more likely to illustrate variety rather than typicality.

In practice, however, the ideal of providing as many examples as possible has to be balanced against space constraints, potentially necessitating a trade-off between number and length of the examples provided. On this point, Fox (1987, p. 149) argues

for the provision of few but long examples in the name of authenticity:[4] "Anything really typical is space-consuming; and yet to give shortened versions is misleading. It is therefore perhaps preferable to give one or two longer examples as opposed to four or five shorter ones." The cultural distance between L1 and L2 of the language learner may also play a role — Jacobsen et al. (1991, p. 2788) point out that "if the non-linguistic background of the two languages is largely the same, fewer examples are needed".

2.5 Sources of Example Sentences: The 'Invented' *vs* 'Authentic' Example Debate

With regard to their respective sources, three types of examples can be identified. They may be citations taken directly from a corpus without any modification by the lexicographer (corpus-attested or 'authentic' examples); they may be corpus citations that are in some way modified by the lexicographer, usually by shortening them in order to simplify vocabulary or grammatical structures (corpus-oriented examples); or they may be concocted by the lexicographer based on his or her intuitions about the target word in question ('invented' examples). Zöfgen (1994, p. 156) observes that the metalexicographic discussion on examples tends to be confined to a debate on the respective merits on invented and quoted examples; it certainly appears to be the most intensely debated and controversial issue among L2 lexicographers.

2.5.1 Corpus-attested examples

Many authors extol the virtues of corpus-attested examples mainly on the grounds of their perceived *authenticity*, *representativeness* and *naturalness* (Fox, 1987). The quality of being authentic is often considered the primary criterion for the quality of an example; Abel (2000) notes that this is true in particular for monolingual dictionaries not primarily geared at L2 learners. Zgusta (1971, p. 265) opines that a corpus-attested example has "the great advantage that it has a highly factual character; evidence can be produced that a word in question really was used in a certain passage by a certain author". Arguably the strongest advocates of corpus-attested examples are lexicographers involved in the *COLLINS COBUILD* lexicographic books and resources designed for learners of L2 English.

[4] see also the discussion on invented *vs* authentic examples below.

Fox goes farther than most other lexicographers when she claims that "authentic examples are almost always superior to made-up ones" because "real examples have actually occurred in the language" (Fox, 1987, p. 149). According to Fox, it is of the utmost importance for examples to show how the target words are typically used, while conceding that "it is no easy matter to find examples that are typical" (p. 139). This caveat is echoed more strongly by Cook who cautions that "to establish typicality, comparison of a large number of attested examples is needed" (Cook, 2001, p. 377). Fox acknowledges that corpus examples may contain difficult vocabulary but insists that "it is better to give a slightly difficult example [...] than to give one that has been made up and does not sound natural in all its details" (p. 146). For Fox, the fact that quoted examples are necessarily taken out of their original context and therefore have 'loose ends' is a virtue rather than a liability. She believes that isolated, made-up examples lacking these loose ends may be too elaborately spelled-out, thus discouraging learners from thinking about the meaning a word. Sinclair (1987) believes that quoted examples are superior to invented ones because they could fulfill the dual function of encoding and decoding, i.e. they could serve as models for both production and reception, while made-up examples could only be used for the latter purpose.

Fox's extreme position that "authentic examples are almost always superior to made-up ones" appears questionable, especially as it is not backed up with empirical evidence, and in fact has been criticized on various grounds. Bogaards (1996, p. 309) observes that many examples in *COBUILD* are rather long and contain infrequent words, which he considers problematic even for production purposes: "The problem with these examples could be that because of their length as well as of the presence of unfamiliar elements, they do not present in a clear way the structure that was to be illustrated and they cannot easily be taken as models for the learner's own production".

Cook (2001) argues that the difference in authenticity between quoted and made-up examples that Fox and Sinclair appear to take for granted is in fact circumstantial rather than linguistic: he points out that "something which was authentic when used is no longer authentic when repeated for pedagogic purposes" , and that "conversely an IS [invented sentence] can easily become authentic if it is used for some non-pedagogic purpose — say for the beginning of a story" (Cook, 2001, p. 378). By a similar token, Cook challenges the straightforwardness of the dichotomy between 'invented' and 'attested' examples: "The utterances in attested data have also been invented, though for communication rather than illustration. The difference is one of purpose." (Cook,

2001, p. 376).

2.5.2 Corpus-oriented examples

There are two possible kinds of compromise between the extremes of invented and authentic examples: corpus examples that are modified by the lexicographer (usually by removing difficult vocabulary, or information that is considered distracting and non-essential), and constructed examples that are based on corpus evidence. Since the distinction between these two types is often a fine one, they will be considered together for the purposes of this section under the general rubric of 'modified corpus examples'.

Some lexicographers feel that the modification (i.e. simplification) of corpus-attested examples may be advisable for L2 learner dictionaries for two reasons: first, the vocabulary contained in authentic corpus-attested examples may be too difficult to meet the receptive (decoding) needs of the L2 learner. In contrast to native speakers who only need to use the *encoding* function of examples, L2 learners also need to use the *decoding* function, which is arguably of primary importance to them.

Second, corpus examples tend to contain distracting extralinguistic information (such as culture-specific place names and proper nouns) that is too context-specific, as well as discourse features such as deictic expressions[5]. Xu (2005) investigated the treatment of deictic expressions in five English learners' dictionaries[6] and found that the types of frequencies of deictic expressions in example sentences largely correspond to those in normal discourse.

Cowie argues that stripping away this information may even be more helpful for production than a lengthy example "which provides superfluous detail in the name of authenticity" (Cowie, 1999, p. 137). On the other hand, Schouten-van Parreren (1989, p. 80) cautions that the lack of redundancy of simplified texts may be problematic, as "an apparently easy, adapted text may be more difficult to understand than its authentic counterpart."

Several authors appear to advocate the use of simplified, but still corpus-oriented, examples: Jacobsen et al. (1991, p. 2788) observe that "often an authentic example can be abbreviated or paraphrased without losing any of its illustrative value". Striking a similar note, Zgusta (1971, p. 265) writes that "probably the best thing to do is quote [...] a reduced part of a passage in a text from which those parts that are inessential

[5] e.g. *this, that, you*

[6] While the investigated dictionaries were described as corpus-based, it appears to be unclear to what extent they used modified citations.

are omitted". However, Cowie (1989) cautions that lexicographers may be tempted to make their examples too concise and thus render them artificial. As a case in point, he cites the following example for the target word *operational*: *"When will the newly designed aircraft be operational?"*, which, Cowie maintains, would be more naturally expressed as *"It's a newly designed aircraft." "When will it be operational?"*.

2.5.3 Invented Examples

Despite the current availability of electronic corpora, lexicographers still use invented sentences to exemplify target words, either because no suitable quotations are available for the target word in question, or because of a deliberate decision to use invented examples for pedagogical reasons.

Examples invented by lexicographers are rejected by the proponents of authentic examples such as Fox and Sinclair who claim that they are often stilted, unnatural, and fail to show the target word in *typical* contexts. Fox (1987) argues that invented examples are necessarily unnatural as they lack the "loose ends" of quoted examples, thereby giving the false impression that language is a series of isolated sentences. Fox's rejection of any kind of invented examples is pithily summarized by her claim that "we cannot trust native speakers to invent sentences except in a proper communicative context" (Fox, 1987, p. 144).

Laufer (1992) and Cook (2001), among others, have argued that, for the most part, these objections do not stand up to scrutiny. Laufer (1992, p. 72) points out that there seems to be no *a priori* reason to believe that lexicographers as educated native speakers of the language cannot have correct intuitions about the typical usage of a word or its typical linguistic environment. But even if the claim that invented examples are less natural was correct, she proceeds to argue, "we might still prefer to see them in learner's dictionaries if their pedagogic value proved to be greater than that of the authentic ones" (Laufer, 1992, p. 72). Striking a similar note, Cook (2001, p. 377) points out that "the inventor of an IS [invented example] may have other criteria than realism. The intention may even be to skew or simplify the language deliberately for some pedagogic reason." Invented examples, Cook believes, are a means of promoting noticing precisely *because* they are isolated and decontextualized; to him, neither situation nor co-text are necessary prerequisites to processing. Empirical studies by Maingay and Rundell (1990) and Laufer (1992) cast further doubt on the notion that quoted examples are necessarily more natural than made-up ones. Maingay and Rundell (1990)

found that teachers of L2 English in general proved unable to tell one type from the other, while Laufer (1992, p. 73) reported that "there was no correlation between the source of an example and its perceived pedagogic value".[7]

Laufer (1992) compared the pedagogical value of corpus-attested and invented examples in terms of the difference in learner performance on new words (for advanced learners of L2 English and learners' dictionaries). She reported that "lexicographer's examples are more helpful in comprehension of new words than the authentic ones. In production of the new word, lexicographer's examples are also more helpful, but not significantly so". She cautioned, however, that "further studies would be useful to substantiate this claim" (Laufer, 1992, p. 76). As a very general guideline for the usage of invented examples, Hermanns (1988) advised that made-up examples are only useful to the extent to which they can evoke a (textual or situational) context in which the example can fulfill its (communicative) function.

2.5.4 Vocabulary used in Examples

An oft-cited criterion for the usefulness of examples is that they are *comprehensible* to the L2 learner: the more vocabulary items used in the example are unfamiliar to the learner, the less likely it is that the example will be helpful to him. According to Drysdale (1987, p. 213), examples are comprehensible if they use styles, registers and vocabulary that are "both idiomatic and intelligible at the students' level of comprehension". Made-up or modified corpus examples can clearly achieve this aim better than authentic corpus examples, as it is difficult to find citations (even in large corpora) which meet these criteria simultaneously (Drysdale, 1987; Zöfgen, 1994). Therefore, many lexicographers and L2 Vocabulary Acquisition researchers propose that the vocabulary for L2 examples should be controlled at least to some extent (Beheydt, 1987; Neubauer, 1989; Abel, 2000), though not necessarily through a limited defining vocabulary, as this may result "in rather stilted or simplified sentences" (Herbst, 1989, p. 1382). Laufer's (1992) study indicates that vocabulary control may be less of a problem for made-up examples compared to authentic ones, as she found the usefulness of the former to be less dependent on the learner's general lexical knowledge. Zöfgen (1994, p. 135) points out the lack of empirical evidence for the assumption that a drastically reduced vocabulary in examples automatically leads to better comprehension.

[7] In Laufer's study, the examples were judged by native speakers who were familiar with the words that were illustrated.

2.6 Functions of Example Sentences

Although there is no generally accepted theory of the lexicographic example as yet (Kempcke, 1992), several attempts have been made to list criteria that examples found in a dictionary should meet. As Harras (1989, p. 608) points out, when considering such a list it is important to bear in mind both the type of dictionary and the target users (in this case, L2 learners). Two key functions of examples mentioned above are sufficiently vague to appear to be universally agreed on in the metalexicographic literature (regardless of the type of target users): to supplement the information in a definition, and to show the target word in context.

It was already mentioned that examples are generally assumed to fulfill the dual function of decoding and encoding. Cowie (1989, p. 67) elaborates on this distinction in the following way:

- Functions relevant to *decoding*:
    1. help to clarify individual meanings;
    2. help the user to distinguish between related meanings.
- Functions relevant to *encoding*:
    1. help the user to select the correct grammatical pattern(s) for a given word or sense;
    2. help the user to form acceptable collocations;
    3. help the user to compose according to native stylistic norms.

Harras's (1989) list of criteria for quality examples expands on Cowie's classification:

1. the example should demonstrate *prototypical characteristics* of the word;
2. the example meets criterion (1) *and* acts as an *implicit* definition (or "forcing example") where the surrounding context determines the word meaning to such an extent that a meaning paraphrase is unnecessary (see also Zöfgen (1994, p. 188))[8];

[8]Strong determination (forcing example): "*Service is included in the check so there is usually no* **tipping** *in Germany.*" *vs* weak determination: "*I don't like* **tipping***.*"

3. the example meets criterion (1) and is an *authentic* example;
4. the example contains sense-related words (e.g. synonyms, antonyms) of the word;[9]
5. the example displays a characteristic aspect of *common attitudes* towards the word;
6. the example displays a particular *manner of speech (parlance)* typical of the text domain in which the word is characteristically used;[10]
7. the example is *meta-communicative* and documents evaluations of the word's value in usage;[11]
8. the example acts as an *exemplum in contrario*, i.e. documents unusual and creative word usages.[12]

It should be noted that an example cannot meet all of these criteria simultaneously.[13] Harras herself acknowledges several shortcomings of this list: the weighting of these maxims is in need of empirical validation, and the criteria have been developed for a 'traditional' general-purpose dictionary, rather than an L2 learner dictionary. It is apparent that some of Harris's criteria may not always be applicable to learner dictionaries (e.g. criterion (6) may only be achieved at the cost of comprehensibility, as demonstrated by the example).

Looking at Harris's list with the target group of L2 learners in mind, the criteria seem to be listed roughly in the order of descending importance. (Proto)typicality and naturalness of examples are widely agreed upon maxims for examples aimed at L2 learners. The more implicit information an example conveys (*forcing examples*), the more potential it has to be useful to L2 learners (Zöfgen, 1986, 1994). However, Zöfgen cautions that strong context determination should not come at the expense of the example's naturalness (in the sense of placing the learner in a typical communication situation). As the discussion above suggests, authenticity *per se* (in the sense of the examples being unmodified corpus citations) is arguably less important for L2

[9] e.g. *Far from being* **ugly***, she was the prettiest girl I could have hoped to meet.*

[10] e.g. *In Deutschland wurde das Projekt* **Moderne** *von Hitler gestoppt.* (In Germany, the project 'modern age' was stopped by Hitler.)

[11] e.g. *Es ist nicht nur das Stilverlangen der* **Moderne** *zerbrochen, jener ästhetisierende Reinigungsfanatismus, dem nicht nur das aus der Antike überkommene Ornament, sondern die Tradition selber Verbrechen war.*

[12] e.g. *It was a bright cold day in April, and the* **clocks** *were striking thirteen.* (George Orwell, *Nineteen Eighty-Four*)

[13] e.g. compare (1) *vs* (8); (2) *vs* (3).

learners than it is for native speakers. Criteria (5)-(7) can arguably be subsumed under Cowie's third encoding function (helping the user to compose according to native stylistic norms). Finally, the usefulness of (7) and (8) is doubtful, as these criteria apparently contradict the above maxims of naturalness and typicality.

Notable through their absence from Harras's list are two criteria that are of obvious importance to L2 learners: ensuring *comprehensibility* by using familiar vocabulary, and — to a lesser extent — relating examples to the world knowledge of language learners (De Florio-Hansen, 1994).

2.7 Conclusion

The above discussion shows that there is a wide variety of opinion on what form the examples should take, what functions they should fulfill and, in particular, what their sources should be. Most of the metalexicographic debate has been shown to center on the respective merits of invented *vs* authentic corpus examples; the potential compromises of either adopting simplified authentic examples, or using examples that are invented by the lexicographer using corpus statistics as a guideline (corpus-oriented examples), are also controversial. This lack of consensus is reflected in current learner dictionaries for L2 German, which for the most part do not provide any illumination in their prefaces as to the way their examples were derived. A partial exception to this is the current monolingual L2 learner dictionary *Großwörterbuch Deutsch als Fremdsprache* (Langenscheidt, 2003), which states that for didactic reasons it also uses examples that are not corpus-attested, albeit without providing further elaboration.

If a general trend in both the L2 metalexicographic discussion and the current lexicographic practice as reflected in L2 learner dictionaries is discernible, it is arguably the following: L2 lexicography seems to have moved beyond the exclusively corpus-fixated position epitomized by *COLLINS COBUILD* lexicographers in the 1980s (Sinclair, 1987; Fox, 1987). It seems to have been replaced a more flexible approach that — as exemplified by the above-mentioned Langenscheidt (2003) — takes the insights of corpus linguistics on board, yet is not bound by them if invented examples appear to be of greater pedagogic value. This policy directly reflects the position taken by authors such as Cowie (1983), Drysdale (1987), Moulin (1983), and Zöfgen (1994) who all suggest that examples should be corpus-*oriented* whenever possible, but should be chosen *primarily* on the basis of their comprehensibility and usefulness for the learner, rather than merely on the grounds of authenticity and corpus-attested frequency.

To summarize the survey of the L2 lexicographic literature on example sentences, there is general disagreement on functions, forms, and sources of example sentences in learner dictionaries. Insofar as specific claims have been made (Fox's assertion that authentic examples are almost always superior to invented ones), they appear questionable and lack empirical validation. The remainder of this thesis will therefore address the issue from a pedagogic, teacher-centered point of view. Given that the scant evidence available on the pedagogic usefulness of invented *vs* authentic examples suggests that the former are of greater pedagogic value at least as far as comprehension is concerned (Laufer, 1992), it appears justified to use the invented examples of experienced L2 teachers as a yardstick against which corpus examples can subsequently be measured. The prerequisite empirical study that was conducted with experienced teachers of L2 German will be discussed in the following chapter.

Chapter 3

An Exploratory Study With Teachers of L2 German

3.1 Introduction

In chapter 2, an overview has been provided of the discussion on example sentences (ES) in the lexicographic and Second Language Acquisition literature, as well as on the actual use of ES in modern learner dictionaries. The general picture that has emerged from this overview is that the lexicographic debate on ES has largely focused on the question of whether to use authentic (corpus-*attested*) or corpus-*oriented* examples (which would leave the lexicographer free to invent his own ES where he sees fit). While authentic examples appear to be favored by the vast majority of current native speaker-oriented (monolingual) dictionaries, the situation is much less clear for learner dictionaries, which tend to differ quite considerably in the leeway they give the lexicographer regarding the use of invented examples.

This diversity among learner dictionaries points to the fact that the question of most interest for this study — *What makes a good ES for an L2 learner from an L2 teacher's perspective?* — has either been largely neglected so far in the lexicographic discussion on ES, or has been addressed from a mainly corpus linguistics view lacking empirical evidence. According to this view, an ES is a good example only insofar as it possesses the qualities of *naturalness* and *typicality* (of syntactic usage, collocational patterns etc.) that are generally considered attributes of suitable authentic corpus examples.

However, the discussion in chapter 2 has also hinted at the need to question whether the picture presented above is complete or even fully accurate from a language learner's point of view: not only does the issue of vocabulary simplification need to be consid-

ered, but other factors such as the memorability of the ES (the quality of being "out-of-the-ordinary" in some respect, or interesting by relating to the learner's world of experience), and the concept of *forcing examples* (where the context maximally constrains the meaning of the target word, possibly at the cost of the ES being natural) may play a role as well. The same goes for the presence of lexical items in the ES that are related to the target word (e.g. synonyms). While this is sometimes mentioned in passing as a desirable quality of ES, none of the learner (or even native speaker) dictionaries known to the author make any sort of statement about whether and to what extent they have considered this aspect in their selection of examples.

As has been argued in chapter 2, an empirical approach to the question of what makes an ES useful for the L2 learner would shed some much needed light on the problem of ES selection from an L2 teaching perspective.

Moreover, one aspect relevant to the current study has been left out of the discussion of ES so far. Since the intended future application of this work is a reading environment where learners can click on unknown target words in L2 reading texts, the question of interest for the study at hand can be put in more specific terms: what makes a good ES for an L2 learner *reading an unknown or difficult target word in a given sentence/text?* It could well be hypothesized that an example that relates (i.e. is semantically similar) to the original sentence containing the target word is preferable to a completely unrelated sentence. The goal pursued by the study described in the remainder of the current chapter is to address this question empirically from a *pedagogic* perspective, i.e. that of L2 teachers.

The purpose of this chapter is to describe the study, to analyse it and to discuss its results in terms of the explanations and rationales given by the teachers. Section 3.2 states the purpose of the study in greater detail, while section 3.3 describes the design and method of the study. In section 3.4 the results of the analysis of the teacher explanations are presented, which are then discussed in section 3.5. Section 3.6 provides a summary of the chapter.

3.2 The Purpose of the Study

As has been mentioned in section 3.1, the purpose of the study with L2 teachers is to address the question *"What makes a good ES for an L2 learner reading an unknown or difficult target word in a sentence/text from an L2 teacher's perspective?"*. The study does not claim or strive for definitive answers on this question, but is rather intended

as a first empirical exploration of the issue.

It was mentioned in chapter 2 that ES generally serve a dual purpose — that of *encoding*, i.e. exemplifying the usage, and that of *decoding*, i.e. clarifying the meaning, or distinguishing between related meanings, of the target word. For the current study, it is the *decoding* function that is of interest, i.e. an ES is considered to be a good or helpful ES *if it illustrates the meaning of the target word (as used in the original sentence)*.

The main purpose of the study is twofold: first and foremost, to elicit ES in the form of invented examples from teachers of L2 German for difficult or unknown target words in a given reading text *that they believe to be the most helpful in each case*, and second, to provide an explanation for their choice of ES.

Thus, the main data to be analyzed is twofold: the actual teacher examples, and the explanations the teachers gave for selecting the examples. These data were then used in the following ways: the teacher examples constitute the main basis for the analysis described in the following chapters, while the explanations given by the teachers serve mainly as a guideline to decide which dimensions will be considered in that analysis. Whether or not the chosen analysis dimensions are *significant* criteria of the teachers' choice of examples is a question to be addressed in the analyses of the ES data. It should be noted, however, that the teachers' explanations serve *only* as a guideline, i.e. they have to be corroborated by inspection of the actual examples; by the same token, if eyeballing the examples reveals obvious analysis dimensions that teachers have failed to mention explicitly or have only mentioned rarely, then these dimensions may suggest or strengthen criteria not obvious from the explanations. It is only the analysis of the teachers' *explanations*, and the implications they have for the main analysis of the ES data, that are discussed in this chapter (sections 3.4 and 3.5).

3.3 The Design of the Study

The study has been designed as a hard-copy questionnaire asking each participant to complete three main tasks in order: (i) identify difficult target words in a German reading text; (ii) provide a made-up example deemed maximally helpful by the participant; (iii) explain the respective choice of ES. Given that task (i) is essentially just a prerequisite for carrying out tasks (ii) and (iii), which provide the main data of interest for the study, and that completing the questionnaire required the participants to dedicate a considerable amount of time and effort to the task (around one hour), the study has

been designed in a flexible way.

In concrete terms, this meant that participants had the option of not fully completing task (i) if time constraints seemed to militate against it, and to rather concentrate on the provision (and explanation) of ES. It also meant that, even though the presence of the experimenter was deemed advisable in order to be able to answer any questions or clarify issues that may have remained unclear even after reading the detailed instructions, participants had the option of completing the questionnaire in their own time if they offered to do so after the one-hour session with the experimenter.[1]

3.3.1 Participants

Altogether 17 participants took part in the study. All of the participants were teachers of L2 German at German departments of universities in Scotland or language teaching institutions such as the Goethe-Institut, with at least three years of teaching experience.[2] All teachers were teaching German at an advanced level of proficiency, i.e. their target learners had attained a sufficiently advanced level of L2 German where syntax does not present an insurmountable obstacle to reading and text comprehension. In terms of the Common Reference Levels given by the Council of Europe[3], all teachers had students at least at the B1 level of both General Language Proficiency and Reading Proficiency[4]; about half of the participating teachers (nine) had students in the C (C1 or C2) categories of reading proficiency.

The L1 of the participants' students was predominantly English. 13 of the teachers were native German speakers (one was a native speaker of Austrian German), the rest (four) were native speakers of (British) English. Participation in the study was voluntary and unpaid.

[1] Two of the seventeen participants took advantage of this opportunity.

[2] The exception being two teachers whose teaching experience was only one year respectively two terms.

[3] These can be briefly summarized as follows with respect to comprehension: (for General Language Proficiency/Reading Proficiency respectively): A1: 'Can understand and use familiar everyday expressions and very basic phrases'/'Can understand familiar names, words and very simple sentences'; A2: 'Can understand sentences and frequently used expressions related to areas of most immediate relevance'/'Can read very short, simple texts'; B1: 'Can understand the main points of clear standard input on familiar matters...Can deal with most situations likely to arise whilst travelling...'/'Can understand texts that consist mainly of high frequency language'; B2: 'Can understand the main ideas of complex text'/'Can understand articles and reports concerned with contemporary problems'; C1: 'Can understand a wide range of demanding, longer texts, and recognise implicit meaning'/'Can understand long and complex factual and literary texts'; C2: 'Can understand with ease virtually everything heard or read'/'Can read with ease virtually all forms of the written language'

[4] The exception being one teacher who only had students she deemed to be at the A2 level of General Language Proficiency, but up to B1 level in Reading Proficiency.

3.3.2 Materials

The materials handed out to the participants consisted of two parts: the reading texts where the teachers' task was to identify difficult words, and the questionnaire proper.

The reading texts were taken from contemporary issues of German newspapers and magazines, using both on-line and hardcopy versions. The articles were chosen in such a way as to ensure a broad appeal, i.e. specialist texts of a technical nature, or articles on topics likely to be of a limited or specialist interest only (e.g. gardening, economy reports, technical topics) were avoided. Some of the articles chosen were in the form of editorial commentary or interviews. The topics covered include events in German and international politics that could be assumed to be of general interest, e.g. climate change, as well as travel- and lifestyle-related topics. The articles were typically 1-3 pages in length (ca. 700-1600 words).

The questionnaire handed out to the participating teachers consisted of three parts: (i) the introduction section explaining the tasks, (ii) a section asking each participant to provide some relevant details on his language teaching background, and (iii) the actual questionnaire form.[5]

The introduction section was split into two tasks: (1) the summary of the three tasks given above, which was provided before the background section; (2) a detailed description of the tasks after the participants had completed the background section.

The actual questionnaire form provided identical templates for 40 target words, each consisting of the following parts: spaces for entering the target word, the corresponding example sentence, and several lines for the explanation and criteria for the choice of the example. At the top of the form, an example for an ES was provided for the hypothetical target word *Trinkgeld* (tip): *In diesem Restaurant ist der Service im Preis inbegriffen, aber es ist trotzdem üblich, dem Kellner ein* **Trinkgeld** *zu geben.* (In this restaurant, service is included in the price, but it is still usual to give a tip to the waiter). A sample questionnaire is provided in full[6] in Appendix A.

[5] In the following, only the parts of the questionnaire that have been used for the analysis are mentioned. The remainder are data that are not of direct relevance for the study at hand, but might be of interest for follow-up studies; they include the following: (a) information provided by the teacher as to which of the classes he was teaching (the proficiency level of which he provided in the background section); (b) classification of the unknown target word as regards the reason why the word is likely to be unfamiliar to the students. Information provided in the background section was also not used in the analysis but is summarized below.

[6] The appendix only provides the first 3 target words of section (iii) of the questionnaire, as all following pages are identical except for the target word numbering.

3.3.3 Procedure

In the introduction section of the questionnaire, participants were first asked to read the text provided, and then to identify all words[7] that they believed to cause difficulty (i.e. words that they thought might be unknown to some degree) to the average learner in their selected group of students. The teachers were free in their selection of student group for this task, but were told that a class nearest B2 level would be preferable. They were then asked, for each target word, to give the best example sentence they could think of to *illustrate the meaning* of the word, taking into account: (a) the proficiency level of their students; (b) the surrounding context of the target word in the reading text; and (c) general interest areas and the world knowledge of their students. Teachers were also told that they should provide *exactly one* sentence for each target word, that their example should contain the target word in question, and be distinct from other explanation options such as definitions or paraphrases. Finally, teachers were asked to provide, for each example sentence and in as much detail as possible, an explanation as to why they considered that sentence to be particularly helpful for illustrating the meaning of the word, and to list any criteria they used in their choice of the ES.

In the background section, teachers were asked to provide the following information: the length of their teaching experience of L2 German, any other languages they were teaching apart from German (and for how long), their native language, and their practice of using example sentences in the classroom (on a 5-point scale from 'never' to 'always'). For each of the classes they were teaching, the participants were also asked to provide the class title, the number of students in each class (broken down by their respective L1), and the General Language Proficiency and Reading Proficiency Levels according to the Common Reference Levels of the Council of Europe included in the section.

3.4 The Results

3.4.1 Pre-screening

After screening the questionnaires returned by the participants, a total of 243 example sentences were retained for the analysis as valid examples. The number of valid ex-

[7] In the instructions, teachers were told to not only consider single words but also multi-word lexical units such as special collocations or idioms, e.g. *ins Gras beißen* (to kick the bucket). As mentioned in section 3.4, however, it was subsequently decided to restrict the analysis to single words only.

amples varied quite considerably from teacher to teacher (lowest number: 1, highest number: 23). The 243 ES correspond to 240 unique target words as 3 words are shared by two ES.

ES were discarded after the screening process due to one of the following reasons:

- *Multi-word lexical items*: the target word identified by the teacher was not a single word, but a multi-world lexical item that, in some of its forms, occurs as discontinuous constituents.[8] The decision to eliminate multi-word lexical items was made mainly for practical reasons, in order to facilitate both the eventual selection of corpus examples and some of the subsequent analyses (in particular the analysis of significant co-occurrences). Multi-world lexical items that were eliminated include the following subgroups:
  - *Idioms* such as *ins Gras beißen*;
  - *Verbs with particles*, in case they are semantically distinct from the same verb forms used without a particle (e.g. *verfallen auf* (to hit upon [an idea]) vs *verfallen* (to expire)). In cases where the particle could be left out without a change of meaning of the verb (e.g. *glauben an* (to believe in)), the target word was modified accordingly.
  - *Reflexive verbs*, in case the reflexive verb has a semantically distinct, non-reflexive version. Example: Target word *sich verschreiben* (to commit oneself to) vs *verschreiben* (to prescribe).
  - *Verbs with separable prefixes or multi-constituent verbs* (e.g. *abnehmen* (to believe) or *satt haben* (to have enough of something)[9].
- *Definitions:* Despite the experimenter and the instructions emphasizing the use of ES *as distinct from* definitions, several ES submitted by the participants were either straight definitions or explanations with or without using the target word (along the lines of 'X is a/used for/made of/etc. Y') or very close to definitions. The decision of when an ES was considered too close to a definition was based on introspection of the experimenter, the crucial criterion being the "naturalness"

[8] The latter qualification means that certain multi-word adverbs or fixed expressions, which always occur as continuous constituents, e.g. *nach Belieben, gesetzt den Fall*, were retained for the analysis.

[9] Verbs with separable prefixes were eliminated even if they appeared in their unseparated form in the ES, since for some of the subsequent analyses (e.g. analysis of significant co-occurrences, see chapter 6) and the final selection of corpus examples, the entire lemma (including the separated forms) would have to be considered as well.

test, which can be paraphrased by the question 'Could this ES have appeared in a newspaper/magazine text that does not strive to explain or define the target word in question?'. It should be noted that this decision, while reasonably straightforward in most cases, was not always a clear-cut one: for instance, the ES for the target word *Teig* (dough) *Brot wird aus Teig gebacken, der sich aus Mehl, Zucker, Salz, Hefe und Wasser zusammensetzt* (Bread is made of dough, which consists of flour, sugar, salt, yeast and water), though technically appearing to be a definition for *Brot* (bread) rather than *Teig*, still seemed too close to a definition for the latter as it does not pass the 'naturalness' test mentioned above. On the other hand, an ES (for the target word *Landesbediensteter*) such as *Jeder Beamter ist gewissermaßen ein Landesbediensteter* (Every public servant is, as it were, a civil servant employed by a *Land*), while also similar to a definition, was accepted as an ES not only because the definitional aspect seems to be more "loose" than the one in the previous example, but more importantly because the ES is conceivable to appear in certain contexts in newspaper or magazine articles (e.g. on the structure of civil service etc.).

- *Target word not used in ES:* This concerns mostly cases where a word form is used in the ES that belongs to lemma which is morphologically closely related but distinct from the lemma of the target word (for example: nominalizations of verb forms such as *Wucherung* (proliferation) vs *wuchern* (to proliferate)). A second (less frequent) case concerns ES where the target word is paraphrased by a synonymous word.

- *More than one ES:* In these cases, the ES proper containing the target word was either followed or preceded by (one or more) context sentences *that could not straightforwardly be combined into one sentence by replacing the sentence boundary "." by 'non-boundary' markers such as a colon or semi-colon* (see also modifications below). This means that combining the sentences into one composite sentence would have seemed stylistically awkward or impossible, e.g. because the ES is preceded by a context sentence ended by a question mark. However, the ES was retained if the additional sentence was not a context, but a definitional or explanatory sentence, in which case the additional sentence was disregarded. Example of an excluded ES (Target word: *krachen* (to crash)): *Was ist das für ein komisches Geräusch? Es hört sich an, als ob ein Auto in eine Mauer gekracht ist.* (What kind of strange noise is that? It sounds like a car

crashed into a wall.)

In the following cases, ES (or their corresponding target words) were retained but modified:

- *Compound target words* that were correctly used in the ES but shortened to one constituent in the target word as identified by the teacher (e.g. *EDV-Fachmann* (computer expert) shortened to *EDV* (electronic data processing)). In these cases, the target word entries were corrected accordingly.

- *Multi-word target words* where shortened to single words were this seemed possible or appropriate. This includes the following cases:

  - *Reflexive verbs* such as *sich rasieren* (to shave) were stripped of their reflexive affix so as to conform to the exclusion of multi-words rule (see above). The same goes for verbs with predicates, unless the predicate indicates a different meaning of the verb (see above).

  - *Idioms* or *fixed expressions* that seemed more plausible as lexical entries when broken down into (some of) their constituents. Example: *strafverschärfend ins Gewicht fallen* (roughly: leading to an increased severity of the sentence) was broken down into the adjective/adverb *strafverschärfend* (increasing the severity of the sentence), and the fixed expression *ins Gewicht fallen* (to matter, be crucial), the latter of which was eliminated due to the exclusion of multi-word expressions.

- *ES with 'dummy' variables:* This concerned cases[10] where one word in the ES was left open. The missing word was denoted by ..., or a variable such as X, to indicate that any word of an approriate word class could be inserted. Such ES were retained (with a suitable instance of the word class inserted) as long as the word class in question was reasonably narrow and could be unambiguously gleaned from the surrounding context. For example, 'tuberculosis' instantiated a generic illness in the ES (for the target word *rückläufig* (declining): *Früher erkrankten viele Menschen an ..., zum Glück sind die Zahlen seit kurzem rückläufig.* (In past ages, many people fell ill with ...., luckily the numbers have been declining recently).

[10] only one case in practice

- *More than one ES:* This includes cases where the ES proper containing the target word was either followed or preceded by (one or more) context sentences *that could be combined into one sentence by replacing the sentence boundary "." by 'non-boundary' markers such as a colon or semi-colon* (cf. also the corresponding entry in the list of eliminated ES above).[11]

In summary, of the total of 359 ES submitted by the participants, 243 (ca. 68%) were retained for final analysis; of the remainder, 51 (ca. 14%) were not valid ES due to one of the reasons given above, and 65 (ca. 18%) were discarded due to their status as multi-word target words. The full set of all retained ES and their corresponding OS is given in Appendix B.

3.4.2 Results of Explanation Analysis

As has been mentioned above, the analysis results described in this section pertain exclusively to the explanation section of the teacher data; the actual ES only come into play insofar as they serve to confirm the teachers' explanation, or help to disambiguate among different possible interpretations of the explanation. This also means that an explanation provided by a teacher is discarded from the analysis if it is contradicted by the inspection of its corresponding ES.

Before an overview of the teachers' explanations is presented, the following subsection provides a summary of the criteria of whether or not an explanation is discarded or retained for analysis.

3.4.2.1 Criteria for Discarding Explanations

As a general rule, an explanation provided by a teacher is not considered for the analysis if it (a) is not associated with a 'valid' ES (according to the the criteria described above), or (b) is neither sufficiently useful nor informative for one of the following reasons:

- The explanation refers to the *difficulty* of providing a useful example. Examples of this type of explanation are: *"This is a tricky one, because it seems to me that the word is used wrongly in the text [...]"; "The use of this word as a description of eating is clear from the context, however the precise nature of the action is*

[11] The decision of whether to use a colon or a semi-colon as a replacement punctuation mark was based on introspection as to the best stylistic fit by the experimenter.

difficult to convey[...]"; "A slightly unusual use of 'mürbe' here - I'd have to make this clear in any explanation".

- Explanations that are too *specific* or *idiosyncratic* to be generalized to other cases, e.g. *"Trying here to suggest that 'urkundlich' is referring to something official, and written, which can be used as historical proof of an event."; "Suggesting that a case is only official when it has been written down and entered into the bureaucratic system."; "Students will be able to deduce the meaning because of the logic of someone wishing to avoid the sun every day."*

- Explanations that are *unclear* or *too general* to contain any useful information, e.g. *"Hopefully the meaning would emerge from the context"; "fairly easy to understand"; "bit of a long shot, but the association might work for some students"; "semantically supportive sentence context [...]"; "Trying to use the word in a context which helps suggest its meaning [...]"*.

In summary, of the 243 explanations associated with valid ES, 26 (11% of the total) have been discarded for one of the above reasons, leaving a total of 217 explanations to be summarized below.

3.4.2.2 Summary of Results

The results of the analysis of teachers' explanations are summarized in table 3.1. Since the explanation section of the questionnaire did not involve choosing among any predefined labels or categories, the teachers were left free in their wording of any given explanation. Thus the labels provided in table 3.1 are an attempt to summarize the essence of the explanations; the degree to which they happen to coincide with the actual wording of an explanation is obviously variable. It should also be noted that some of the teachers' explanations provide entries for more than one category (i.e. the categories are not mutually exclusive), so that the overall group percentages may add up to more than 100%; also, some sub-categories are listed under more than heading if appropriate.[12]

As can be seen from the table, the teachers' explanations can be grouped into the following main categories (in the order of frequency of mention): context-related, spe-

[12]For example, the subcategory of *similar context* is listed under both "Context (Level Unspecified)" and "Context (Sentence Level)", depending on whether the context is similar on a local (sentence) or global (text) level, and a lexical association of the target word in the ES may also be a frequent co-occurrence.

Table 3.1: Categories of Teacher Explanations (# = number of mentions out of 243)

CATEGORY	#	CATEGORY *(continued)*	#
CONTEXT (TOTAL)	**106**	**USAGE**	**42**
CONTEXT (LEVEL UNSPECIFIED)	**103**	**Similar Usage**	**26**
Similar Context	**16**	same/similar usage	25
similar context	13	keep figurative usage	1
similar context but clarified	3	**Different Usage**	**16**
Different Context	**87**	figurative → literal usage	13
different/particular context	32	different usage (other)	3
more familiar context	30	**PHRASAL CHOICES**	**23**
more constraining context (less ambiguous)	20	use of phrasal association/connotation	10
simpler/more concrete context	8	contrast with phrasal antonym/synonym	8
more typical context	5	phrase to describe targ. word's function	3
more detailed context	1	contrast between two phrases	1
CONTEXT (SENTENCE LEVEL)	**73**	use phrase to show target word is transitive verb	1
Similar Sentence	**5**	**REDUCE SYNTAX COMPLEXITY**	**11**
similar sentence/context	4	**MORPHOLOGICAL CHOICES**	**11**
clarify context	1	**Verbs**	**7**
Dissimilar Sentence	**68**	simpler form (e.g. not subjunctive)	2
more familiar context	27	regular tense (for irregular verb)	2
different/particular context	32	use most frequently used tense	1
more typical context	5	use tense to indicate non-sep. prefix	1
explain function of target word	2	use specific person	1
use historical context	2	**Nouns** (use sing./art. to show gender)	**3**
SPECIFIC LEXICAL CHOICES	**75**	**Adjectives/Adverbs**	**1**
Lexical Relations	**60**	**L1-RELATED (on sentence level)**	**3**
Lexical Associations	39	show difference to L1 false friend	2
Antonyms	11	association with L1 idiom	1
Synonyms	8	**OTHER EXPLANATIONS**	**28**
Hyponyms/Hypernyms	1	quasi-definition couched as example	7
Related word with same root	1	exploit general world knowledge	5
Use of 'context' words	**8**	convey register/style of word	4
Frequent Co-Occurrences	**5**	use of redundancy	3
Use of Cognates	**3**	explain function of target word	2
Use of familiar words	**3**	give reason(s) for target word	2
Use of function words	**2**	use of humor (funny example)	2
		others (single mention each)	3

cific lexical choices, usage, phrasal choices, syntax (i.e. the reduction of syntactic complexity), morphology, L1-related choices, and 'other explanations' that are not covered by any of the main categories above.

Turning first to the most frequently referenced main category in the explanations referring in some way to the **context** of the original occurrence of the respective target word, two subgroups can be distinguished: the first references the original context in a generic sense, i.e. the explanation typically mentions the 'context' without specifying whether it is the context of the piece of text (article or interview), or the narrower context of just the *sentence* containing the target word, that is being referred to. The second group refers to that narrower sentence context alone.[13]

With respect to the 'unspecified context' subgroup, two opposing goals can be identified that can loosely be dubbed 'keep similar context' and 'change context'. In the former, the context of the original text is kept as a deliberate reference point, i.e. the teacher's ES is deliberately specific to the original text, sometimes with the added intention of 'clarifying' it.[14] In the second 'change context' subgroup, the apparent goal of the ES was to change the original context either in a non-specified way (*choose different or particular context*), or with an explicit goal in mind: e.g. to choose a context more familiar to the students given their likely experiences and world knowledge; to constrain the meaning of the target word in the ES, i.e. make it less ambiguous[15]; to either simplify the context (by e.g. making an abstract context more concrete) or elaborate it (by providing more details); or to provide a more typical context for the target word.

Turning to the 'sentence similarity' subgroup, the subcategories here are very similar to the one discussed above (for the 'unspecified context' subgroup), with two notable additions: the choice of an ES that (implicitly) explains the function of a target word; and the provision of a historical context for the target word.

The second most frequent main category in teachers' explanations can be loosely labelled **'specific lexical choices'**. This includes the following possibilities: lexical

[13] As the table shows, the term 'sentence' is not necessarily explicitly mentioned in every explanation that is grouped under the *sentence similarity* heading. Sometimes the inspection of the respective original texts and sentences clearly constrains the reference to sentence level, as the context or topic of the teachers' ES is similar to the original sentence but dissimilar to the context or topic of the text as whole. Also, teachers' explanations to the effect of "explain function of target word" clearly refer to the sentence level alone.

[14] The following explanation is an example of 'context clarification': "I'm not sure the surrounding context will help, since level 2 students are often not yet able to consider an L2 text as a whole. I have simply made the context clear."

[15] This corresponds to one of the goals of ES cited in the lexicographic literature known as *forcing examples*, cf. chapter 2.

relations, words that frequently co-occur with the target word in question, choices related to the L1 of the students (use of cognates), and other specific lexical choices that are not subsumed by any of the categories mentioned so far.

Of these subgroups, lexical relations are by far the most frequently mentioned category. Explanations in this category pertain to a specific choice of at least one word in the ES that is either semantically or morphologically related to the target word. About half of the corresponding explanations (ca. 33% of the total number of explanations in this category) pertain to traditional lexical relations such as synonymy, antonymy, and hyponymy/hyperonomy, with antonyms being used surprisingly often compared to synonyms. Even more frequent (ca. 65%) among the specific lexical choices are *lexical associations*, that is words that in some way (and in varying degress of strength of association) reference a common association, connotation or 'prototypical scenario' of the target word. These words usually refer to typical functions or locations of the target word, they may reference what the word typically applies to, or they may refer to combinations of lexical items that together have some association with the target word.[16] A single mention concerns morphologically related words, in this case the use of a related word with the same root (*Erfinder* (inventor) for the target word *findig* (resourceful)).

The next most frequent subgroup (by quite some distance) of specific lexical choices concern 'context' words; 'context' here is interpreted as referring to a general setting or scenario in which the target can often be found (i.e. this may also be classified as 'co-occurrences' or 'lexical associations', but has not been identified as either in the teacher's comment). Typical teacher comments here are along the lines of 'explanation via context words', or 'Word X provides context for target word'; example: the use of *Zeichnung* (drawing) for the target word *karikieren* (to caricature).

Almost as frequently mentioned as 'context' words are co-occurrences, i.e. words that frequently co-occur with the target word, or typical collocations. Finally, the use of familiar, well-known words is cited in three cases (the exact nature of the relation between the familiar word and the target word may vary from case to case), as is the use of cognates (e.g. the use of *Missionar* (missionary) for the target word *bekehren* (to convert)) and the use of (one or more) function words suggesting some aspect of the meaning of the target word.

[16]Examples are: "reference to place helps" (referring to the use of *Gefängnis* (prison) in the ES for the TW *Internierte(r)* (internee)), or the use of *Gericht* (court) in the ES for the target word *Klage* (charge, lawsuit) "da mit dieser Assoziation das Verstehen des Wortes erleichtert wird" [because the comprehension of the word is facilitated by this association].

Explanations in the **Usage** category refer to the way the target word is used in the ES compared to the OS: either the usage is kept deliberately similar (62% of all usage-related explanations), or the usage of the target word in the ES differs in some way from that in the OS: in the majority of cases, the target word is used in a literal sense in the ES, whereas it has been used figuratively or metaphorically in the OS.

Explanations in the **Phrasal Choices** category pertain to the choice of lexical chunks, i.e. multi-word lexical items, mostly for the following purposes: as a 'phrasal' association, connotation, antonym, or paraphrase, or to describe the function or the object of the target word.

Syntax, that is, the reduction of syntactic complexity of the ES in relation to the OS, has been mentioned in ca. 5% of all explanations, while another 5% of the explanations are motivated by some consideration of **Morphology**. The purpose that can be gleaned from the latter group of explanations is in most cases either to simplify the morphological form of the target word in the OS (e.g. by choosing a regular tense for a generally irregular verb), or to highlight the most *frequently used* form of a target word (e.g. the most frequently used tense of the verb in the ES). Other motivations are often specific to the respective part-of-speech of the target word and include the use of a particular tense to indicate a non-separable prefix of a verb, and the use of the singular or an article to indicate the gender of a noun.

Other categories not mentioned so far are quite infrequent (less than 10 explanations in each category) and can be taken from table 3.1; the most frequent groups of these are the use of a quasi-definition couched as an ES[17], and the use of general world knowledge of the students (e.g. the reliance on general automotive knowledge in the ES (target word in italics): "Als wir die *Öllache* sahen, wußten wir, dass das Auto Öl verliert." (When we saw the *oil slick*, we knew the car was losing oil).

3.5 Discussion

The results of the explanation analysis presented in section 3.4 have to be taken with a slight degree of caution, for the following reasons: first (and perhaps most obviously), the teachers might not have been able to verbalize certain criteria that they were employing, either because of time constraints, or because they were using them 'subconsciously'. An obvious candidate for an example of the latter category is the

[17]The corresponding ES were not close enough to a definition to have been excluded from the analysis.

reduction of syntactic complexity, which, as inspecting the ES using sentence length as an indicator[18] reveals, is frequently employed by the teachers but relatively rarely mentioned (ca. 5% of total valid explanations). Also, it can be conjectured that the use of frequent co-occurrences is more widespread in the ES data than the explanations would suggest: after all, teachers may have hesitated to verbalize this as an explicit criterion as they would have had to rely on their intuition regarding which words are indeed frequent co-occurrences, an intuition which may not always be infallible and that in any case would have to be confirmed by a corpus analysis.

Second, the criteria for discarding teachers' explanations, while facilitating the analysis in terms of categorization of the explanations, implied decisions that may be considered slightly arbitrary, and that were made by one rater only.

Third, as has already been noted, the explanations given by the teachers were not constrained in their format, so in order to arrive at any summative analysis in section 3.4.2.2, their in some cases quite detailed and lengthy explanations had to be 'condensed' into reasonable labels or categories, which can only be reasonable approximations of the teachers' full explanations.

Fourth, while most of the explanations were quite detailed and in-depth, some of the explanations or terminology used were either too short or too vague and ambiguous to serve as a helpful statement about the intention and rationale behind the example. For example, an explanation such as "clearer context", while stating the context of the OS as a point of reference, does not give any indication as to how the clarification of context was intended to be achieved in the ES (e.g. via simplified vocabulary, changing the topic etc.), nor does it specify if *clarification* of context refers to a reduction of ambiguity in the interpretation of the ES (i.e. a *forcing example*), or to a facilitated comprehension of the sentence as a whole due to difficult vocabulary or topic.

Fifth, the instructions for completing the questionnaire were explicitly encouraging the participants to consider factors such as the surrounding context of the target word in the reading text (as well as proficiency levels, interests and general world knowledge of their students). While this seemed advisable as a reminder for the teachers to not judge the target words in isolation (as they mave have been wont to do considering the time constraints), the possibility cannot be excluded that these instructions may have led to a slight bias in the teachers' answers: instead of merely taking the original context into account as instructed, some teachers may have been inclined to take this

[18] As chapter 4 will show, sentence length is indeed a valid measure of syntactic complexity, at least for the purposes of this study.

as an encouragemnent to provide a *similar*, or more familiar context, in their ES, that is, to let the original context influence the choice of their ES is some way.

Finally, the exclusion of examples pertaining to multi-word target words (see preceding section) may have introduced a slight bias regarding the criteria teachers have employed in their choice of ES. While an inspection of the respective explanations discarded from the analysis indicates that at least some of these examples were motivated in part by morphological considerations, the inspection also suggests that (a) these considerations only seem to apply to the group of separable-prefix verbs, as teachers may want to use a specific verb form that demonstrates the separability of the prefix; (b) none of the other categories (including an overall reduction in syntactic complexity) seem to be affected by the restriction to single words.

These caveats notwithstanding, the results clearly provide some insight into what criteria teachers have been employing when choosing their OS, and can thus provide valuable assistance in the decision on which criteria to analyse in detail. First and foremost, even taking the above caveat into account, the *context* of the original target word (either on sentence or global text level) appears to be the most important consideration for teachers in their choice of ES. Even though the instances where teachers prefer to *change* the context in some way outnumber the cases where they opt to keep the context similar, it is *similarity* of context rather than *dissimilarity* that will be analyzed in chapter 5 as a potential criterion for choice of ES. This is simply because similarity in this case appears to be a reasonable pedagogic goal in itself, as an ES context similar to the original context has the obvious advantage of not confusing the student with a context different from the text he is currently reading. A *dissimilar* context, on the other hand, hardly seems to be worth striving for as an end in itself; rather, it seems to be worthwhile only insofar as it acts as a 'catch-all' term for more specific goals *entailing* a change in context, such as rendering the context more familiar, or more detailed; all of these goals, however, would need their own specific measure, which — for the purposes of this study — cannot be implemented in a straightforward way (for instance, 'more familiar' context would need to incorporate some model on the students' experience and world knowledge, while 'more constraining' or 'more concrete' seem too vague to lend themselves to any obvious measure).

More specifically, it is the notion of *sentence similarity* (rather than sentence-document similarity, i.e. the similarity of the ES to the OS rather than to the entire reading text) that will be analysed in chapter 5. While in principle statistic-based measures exist for the second type of comparison as well, the restriction to sentence level

arises mainly from the practical constraint of the intended application, which involves applying the measures developed to corpus data; these, however, are usually available in the form of single stand-alone sentences.

The choice of specific words in the ES has been shown to be an important criterion for teachers too; this concerns in the main semantical relations in the form of traditional lexical relations (synonymy, antonymy etc.), form-based morphological choices, and considerations of usage in the form of frequent co-occurrences (as has been argued above, the relatively infrequent number of mentions of this criterion in the teachers' explanations should be taken as a lower estimate of the actual use of this criterion). The analysis dimension of specific lexical choices (lexical relations, co-occurrences and morphology) will be discussed in chapter 6.

Finally, *syntax*, i.e. the reduction of syntactic complexity from OS to ES, is an apparently important criterion for teachers and will be analyzed in chapter 4.

The dimensions identified so far as relevant to teachers are in keeping with Van Parreren and Schouten-Van Parreren's (1981) findings that L2 readers act on syntactic, semantic (contextual), and lexical/word form (as well as sometimes stylistic) levels when guessing unknown word meanings in a text.

The *usage* category will not be further analysed, as the most important criterion relating to this category seems to be the transitions from figurative usage in the OS to (mostly) literal or (sometimes) figurative usage in the ES. However, Word Sense Disambiguation (WSD) falls outside the scope of this thesis (and any reliable distinctions between figurative and literal usage would require a particularly fine-grained version of WSD). Thus the distinction between figurative and literal usage of the target word will only be considered for the eventual evaluation of the model to be developed in chapter 7 insofar as corpus examples need to be selected for any given target word. As the explanation analysis strongly suggests, if the target word in the OS is used in a figurative sense, then both figuratively and literally used instances of the word in any potential ES are candidates; however, if the target word in the OS is used in a literal sense, then any sentence where the word is used in a figurative sense seems to be out of the question as a suitable example.[19]

The category of *Phrasal Choices* will likewise not be further analyzed due to the restriction of the target word focus (discussed earlier in this chapter) to single words only. As for the remaining categories in table 3.1 not mentioned so far, they are ei-

[19] As will be discussed in chapter 6, WSD-related decisions such as this will be made manually by the experimenter for the evaluation.

ther too infrequently mentioned in teachers' explanations to merit further analysis, too vague to be operationalizable, or are not easily amenable to analysis because of the lack of straightforwardly implementable measures.

An example of the latter is the case of *forcing examples* discussed in chapter 2, i.e. examples where the meaning of the target word is maximally constrained by the rest of the sentence. While this criterion (usually stated along the lines of 'less ambiguous context') is mentioned relatively frequently in the explanation data (ca. 8%), the fact that the example ES provided at the beginning of the questionnaire was a textbook forcing example for *Trinkgeld* (see section 3.3.2) may have introduced a slight bias and contributed to the relatively high figure. More importantly, it is not clear how any measure of the 'constraining' quality of the ES could be arrived at, and even less how it could be straightforwardly implemented for the intended application of corpus examples. Moreover, as has been stated in chapter 1, the focus of the current study is on the investigation of existing well-known measures for relatively 'straightforward' concepts such as co-occurrence, semantic similarity and syntactic complexity.

Finally, some mention should be made of another potential goal of ES discussed in the lexicographic literature (see chapter 2), namely the simplification of vocabulary. While this goal is not directly mentioned in the teachers' explanations, it crops up indirectly in the guise of 'use of familiar (well-known) words' and possibly also in the *context* section under 'more familiar context' (insofar as the change of context is achieved by means of simpler vocabulary).

Even though it appears reasonable to conjecture that teachers would not use vocabulary in their ES that is at least as difficult as the target word in question, the reduction of lexical difficulty in the ES compared to the OS may not be taken for granted. As the early L2 classroom literature on 'simplified input' has suggested (e.g. Chaudron (1983); Parker and Chaudron (1987)), simplification in the language presented to learners may take the form of either 'restrictive' or 'elaborative' modifications, which in the case of vocabulary may mean the use of a clarifying or elaborating paraphrase at the possible expense of including difficult lexical items. However, the fact that the respective vocabulary proficiency for the classes of each teacher are not always on the same level militates against including the reduction of lexical complexity as an analysis dimension; rather, the concept will enter some of the models to be developed in chapter 7 via the restriction of potential examples to sentences only containing the most important 4000 words for learners of L2 German (according to Langenscheidt (1991)).

3.6 Summary

This chapter presented a design of the study which was used to gain insight into criteria teachers were employing when choosing or inventing ES for difficult or unknown target words. More specifically, the study focussed on the choice of ES for target words in reading texts. The study provided two different kinds of data: the ES themselves, and the explanations provided by the teachers for the choice of their respective ES. The explanations serve as a basis for the analysis of the ES data, according to the criteria identified as important to teachers in the preceding sections.

It is the analysis of the latter data that was discussed in this chapter. The teachers' explanations show that the criteria of context (to be analyzed in the form of sentence similarity), reduction of syntactic complexity, and specific lexical choices such as synonyms, antonyms, frequent co-occurrences and words that are morphologically related to the target word, are all relevant dimensions.

The ES will be analyzed along these dimensions in the following chapters, starting with the reduction of syntactic complexity in chapter 4. The goal of these analyses will be to identify a suitable measure for each of the criteria, and then to analyze the example data to show whether or not the dimensions are indeed significant criteria for the choice of the examples.

Chapter 4

Measuring Syntactic Complexity

4.1 Introduction

We have seen from the analysis of the teachers' explanations in chapter 3 that the reduction of syntactic complexity was a possible criterion that teachers employed in their choice of example sentences. Before the reduction of syntactic complexity in the teachers' examples (compared to the original sentences) can be analyzed, the question of how to measure the concept of syntactic complexity needs to be addressed.

Along with lexical complexity (or vocabulary load), syntactic complexity has been recognised as one of the central factors of text readability (Botel and Granowsky, 1972; Alderson, 2000). Studies such as Wang (1970) have shown syntactic complexity to be a significant determiner of sentence comprehensibility. Even though an underlying assumption of the teacher study (see chapter 3) was that the target learners had attained a sufficiently advanced level of L2 German where syntax does not present an insurmountable obstacle to reading and text comprehension, it is therefore still reasonable to assume that, even for them, a complex sentence construction adds to the difficulty of text comprehension. This is particularly the case considering that the above holds true even for native speakers, and that many reading texts presented to the teachers had been taken from magazines or newspapers that tend to use rather lengthy and complex sentences. Therefore, the hypothesis explored in this chapter is that teachers are inclined to reduce the syntactic complexity of their respective examples compared to the original sentences. As was discussed in the previous chapter, this assumption has been lent further support not only by a cursory inspection of the teacher examples, which for the most part are considerably shorter than the corresponding original sentences, but also by some of the teacher explanations that (in ca. 5% of the cases) refer to reduced

syntactic complexity either directly or indirectly (e.g. 'simplified context' and similar comments).

Of course, it is far from clear whether or not an intuitively appealing, but very crude, measure such as sentence length constitutes an adequate measure of syntactic complexity (henceforth referred to as MSC). While several MSC have been suggested in the linguistic and psycholinguistic literature, they tend to rely on different psycholinguistic assumptions and grammar formalisms (some of which are outdated by now), so no clear consensus has emerged regarding a "standard" MSC that is universally accepted.

A further problem with respect to the MSC that have been put forward is that they tend to assume adult native speakers (mostly of L1 English). From an L2 learning perspective, however, this is less than ideal, since recent psycholinguistic studies of L2 sentence processing in real time have shown that "even highly proficient L2 learners behave differently from native speakers when resolving structural ambiguities or processing syntactic dependencies" (Clahsen and Felser, 2006, p. 30). The general tendency revealed in these studies (cf. Clahsen and Felser (2006) for an overview) is that L2 learners underuse syntactic information while they are guided by lexical-semantic and plausibility information at least to the same extent as native speakers. This observation has given rise to different theoretical accounts of L2 sentence processing, e.g. the declarative *vs* procedural memory distinction posited by researchers such as Ullman (2001) and Paradis (2004) that claims that L2 learners have to mainly rely on the former in their grammatical processing, or the more recent 'Shallow Structure Hypothesis' put forth by Clahsen and Felser (2006) that assumes that the syntactic representations of adult L2 learners during sentence comprehension are shallower and less detailed than those of native speakers.

It should be emphasised at this point that in the following discussion and evaluation of different MSC, no stance is taken in this debate, or on the psycholinguistic adequacy of the respective MSC for L2 learners in general. This is for the following reasons: first, none of the current models or accounts of L2 sentence processing, based as they are on limited empirical evidence on a restricted set of grammatical phenomena such as ambiguous sentences or syntactic dependencies, is developed and detailed enough to provide an operationalizable MSC. Second, it should be born in mind that it is the *teachers' perception* of what constitutes a syntactically complex sentence for their L2 students (rather than some measure of the learners' experienced difficulty of L2 sentence processing) that is the issue to be investigated in this chapter.

Given this situation, the aim of the current chapter is first to establish which of the alternative MSC that have been suggested in the literature best predicts the ratings of syntactic complexity that were provided by linguistically trained native speakers of German.[1] Empirically validated, that measure is then used to analyze whether, for the data at hand, syntactic complexity is indeed a significant criterion that teachers used in the construction of their example sentences.

This chapter is organized as follows: first, an overview of the concept and different alternative MSC that have been suggested in the literature are provided in sections 4.2 and 4.3, respectively. Section 4.4 then describes the empirical study that was carried out with native speakers in order to establish the measure that best correlates with human judgments of syntactic complexity. Section 4.5 presents the analysis of the teacher data in terms of syntactic complexity, while section 4.6 summarizes the chapter.

4.2 The Concept of Syntactic Complexity

As has been mentioned above, while many indicators of syntactic complexity have been suggested in the linguistic and psycholinguistic literature over the years, no generally accepted standard measure has emerged so far. The wide range of alternative measures also reflects the fact that the notion of syntactic complexity is far more difficult to pin down than it may first appear: while a definition along the lines of "A syntactic structure A' is (syntactically) more complex than structure A, if it contains more structural information along a particular dimension" (Brettschneider, 1978) is intuitively appealing, it is also necessarily vague and utlimately not very helpful as it raises the question of how to operationally define the concept of *more structural information*. It seems clear that grammatical constructions that are dense, embedded, or structurally ambiguous have "more structural information" than those that do not; however, this insight does not readily translate into any operational definition. While one could think of several indicators that all contribute to complexity of a parse tree (e.g. number of nodes, depth and branching factor), it is far from clear how these factors can be synthesized into a general complexity measure. For example, Smith (1988, p. 250) lists *amount* (the number of linguistic units in a sentence), *density of structure* (more nonterminal node structures within a phrase), and *ambiguity* (different surface struc-

[1] While it is possible that the teachers' perceptions of syntactic complexity do not always coincide with those of native speakers judging the sentences in a non-educational situation, it nevertheless seems plausible to assume that they are reasonably close.

ture interpretations) as the determinants of surface syntactic complexity, but refrains from an attempt to combine these factors into an overall complexity metric.

The multidimensional nature of syntactic complexity can also be gleaned from the fact that complexity increases not only with the number of elements, but also with the number of links between these elements, and the extent to which these links differ. Different MSC tend to be based on different grammar formalisms, and are thus prone to put different emphases on these dimensions; as a result, they may well differ in the complexity ratings they assign to the same sentence.

But syntactic complexity measures tend to differ not only in their underlying grammar formalism, but also in their general approach and theoretical motivations: some measures strive to explain linguistic data such as grammatical acceptability judgments, while the main focus of others is language production (e.g. L2 writing) or sentence comprehension (readability) . Regarding the latter, Frazier (1988, p. 194) notes that "it seems unlikely that direct measures of complexity[...] will by themselves lead to very refined measures capable of predicting the precise complexity of each portion of a text or reveal the nature and source of differences in processing complexity", and concludes that "[...] ultimately it must be a theory of human language comprehension which will provide (embody) the complexity metric for processing[...]".

4.3 Measures of Syntactic Complexity

Before the MSC that we have considered for this study are described, some mention should be made of some early attempts at MSC that have been excluded from consideration. These include measures that seem to be (a) language-specific to English and cannot straightforwardly be transferred to German; (b) measures that can only be regarded as a vague ad-hoc attempt and lack both a sufficient empirical and theoretical foundation; (c) measures that are based on transformational grammar theory.

A prime example of both (a) and (b) is the Syntactic Complexity Formula (SCF) proposed by Botel and Granowsky (1972), which assigns weightings ranging from 0 to 3 to syntactic structures of English. The authors themselves caution that SCF "should be regarded as a directional effort still requiring further validation" and "should not be considered a precise measuring instrument but rather a device for the identification of syntactic structures that affect readability and for the ranking of these structures in terms of their relative complexity." (Botel and Granowsky, 1972, p. 514).

An example of (c) is the Derivational Theory of Complexity (DTC) first outlined

by Miller and Chomsky (1963). DTC assumes a transformational-generative grammar and basically claims that "the complexity of a sentence is measured by the number of grammatical rules employed in its derivation" (Fodor et al., 1974, p. 320). In other words, DTC views (perceptual) complexity as a function of the transformational distance of the sentence from its base to its surface structure, and predicts a direct correlation between length of the derivation and its parsing complexity (that is, the time it takes to parse the utterance). However, DTC has long been abandoned because of its fundamental shortcomings. For instance, Fodor and Garrett (1967, p. 290) point out that empirical data seem to contradict DTC, as some sentences apparently involving more transformations turned out to be easier to process. They point out that sentences close to their deep structure are generally more richly grammatically elaborated, and suggest that syntactic complexity is also a function of the degree to which the arrangements of elements in surface structure provide clues to the relation of elements in deep structure. Another problem with DTC concerns the conflation of grammatical knowledge and processing behavior: DTC does not distinguish between an (infinite linguistic) *competence* and a (limited observable) *performance*, a distinction typically made in many quarters of modern linguistics and psycholinguistics. Finally, DTC made sense only in the old "transformational grammar" model, since no current generative grammar models involve either derivational rules or transformations.

For these reasons, DTC and other indicators based on transformational sentence analysis — such as the ones used in Wang's (1970) study — have not been considered for this analysis. While some of the following measures that have been considered are also outdated in certain respects (e.g. their respective assumptions about grammar), or do not have any theoretical foundation at all (e.g. sentence length), they compensate for these shortcomings by being intuitively appealing.

In addition to the differences in theoretical motivation outlined in section 4.2, MSC can be classified according to the level of syntactic analysis they draw upon: no analysis at all (sentence length), clause-level analysis (T-Unit length, Coordination Index, Staircase-Measure) or constituent-level analysis ('Yngve'-measure, Non-Terminal-to-Terminal-ratio, Frazier-Measure, EIC and SPLT).

Measures that use clause-level analysis do not require a parse of the sentence, as they only take into account the combination of different clauses that make up the sentence (main clause(s), subordinate clause(s) and apposition(s)). They are thus more coarse-grained than measures that are based on the analysis of the constituent sentence structure as derived from a phrase marker.

Even though parse-based measures can obviously draw upon more detailed syntactic information, they are faced with the problem that even for simple sentences, there may be a wide variety of equally acceptable parse analyses to choose from. Deciding among the possible ways to derive phrase markers is not so much of a problem if one wishes to use only one particular measure for all analyses, as the relative complexity ordering of sentences can be assumed to remain roughly unchanged. However, if the goal is to perform a cross-comparison of the different measures (as is the case here), the choice of a particular parse strategy may influence the relative ordering of measures considerably, as some measures may be affected by a particular parse characteristic while others are not.

The multitude of available phrase markers for any given sentence can be due to the following factors: (a) genuine syntactic ambiguity, e.g. the level of attachment of a PP (*The man saw the woman with a telescope*); (b) the underlying grammar formalism, and (c) whether deep (e.g. with VPs) or flat structures (no VPs) are preferred.

Regarding these factors, the following choices have been made for the analysis at hand:

(a) The most plausible analysis (taking into account semantic considerations) is preferred; in practice, this often means deciding on the most plausible level of attachment for a PP or AdvP;

(b) Only 'proper' trees are allowed as phrase markers, i.e. no crossing of branches is permitted;

(c) A simple phrase structure grammar without transformations, traces, or empty categories has been chosen; nodes include <S, VP, NP, PP, AdjP, AdvP, VG, V, N, P, etc>;

(d) For each sentence, a flat and a deep phrase marker have been constructed, so that each of the parse-based measures yields a 'deep' and a 'flat' value for the sentences analysed.

Regarding (d), the main difference between deep phrase markers and their flat counterparts is that in the former, verbal clause constituents are analysed as either VPs (V" in X-bar-parlance), or V"'s (sentence-level constituents), using Uszkoreit's (1987) GPSG framework as a guideline.[2] The reason for basing the syntactic analy-

[2]The other changes in the tree structure are a direct result of the separate status of verbal constituents, e.g. coordination nodes (e.g. for *und; oder* (and; or)) as siblings of V2/V3s, or AdvP/PP as daughters of V2/V3s.

sis on both deep markers (with VPs) and flat markers (without VPs) is that whether or not German contains a VP-node is a controversial topic in the syntax literature, as "constituent tests of the type traditionally used to justify VP in English fail to deliver decisive results for German" (Grewendorf, 1993, p. 1300). In addition, the pecularities of German clause structure and word order (three main clause types with respect to the position of the finite verb; the existence of discontinuous syntactic constituents and complex verb clusters; a relatively free word order in contrast to e.g. English) tend to complicate an analysis of verbal constituents in terms of VP.

The main considerations in deriving the phrase markers have been plausibility, consistency across the analysed sentences, and practicality (ease of derivation); no claim is made about the adequacy of the derived phrase markers as a syntactic description of German.

The flat phrase markers are based on the parse trees produced by the online UIS parser developed at the University of Zurich (UIS-Parser, 2002). Even though the parser handles a variety of sentence constructions including relative clauses, its coverage is not sufficiently deep to handle certain complex constructions (several adverbs or PPs in a clause, some relative clause constructions etc.). In these cases, the gaps have been filled manually based on partial analyses of the UIS parser, and/or suitable analogous UIS phrase markers of similar constructions.

The deep phrase markers containing VPs have been derived manually on the basis of the UIS output, and have been based on the GPSG framework for German suggested in (Uszkoreit, 1987) whenever possible.[3]

In the following, the selected measures of syntactic complexity are discussed in more detail.

4.3.1 Sentence Length

Arguably the most basic (and crudest) indicator of syntactic complexity is sentence length, as measured by the number of words that a sentence contains. Due to its straightforwardness and simplicity of use — no syntactic analysis whatsoever is needed — (average) sentence length has traditionally been the indicator of choice in the syntactic complexity component of readability formulae (Read, 2000). The underlying assumption is that "syntactically complex authors [...] use longer sentences and more subordinate clauses that reveal more complex structrural relationships" (Beaman, 1984,

[3] Since the GPSG framework is a non-exhaustive fragment of basic rules, some clause constructions are not covered by the rules provided.

p. 45). However, sentence length does not necessarily correlate with the degree of subordination in a sentence, especially when writers rely heavily on coordinated structures and compounding. Furthermore, as Read (2000, p. 72) points out, "there is considerable research which shows that to make sentences easier to understand, words may have to be added, not deleted".

4.3.2 Mean T-unit Length

The T-unit, or Minimal Terminable Unit, is defined as the shortest unit which a sentence can be reduced to, i.e. a main clause including all subordinate clauses and non-clausal structures attached to or embedded in it. For example, the following sentence would count as 2 T-units: [In 1991, my husband got a scholarship from Louisiana Tech University] [and he came to this country to continue his education].

The concept has been introduced by Hunt (1965) as an instrument for measuring the development of syntactic complexity in the writings of schoolchildren. Mean T-unit length does away with one of the most obvious shortcomings of sentence length as a syntactic complexity measure: the failure to assign low complexity values to very long sentences composed of several compounded main clauses. However, the index still fails to deal with excessive coordination within a sentence (Sotillo, 2000).

As Hunt's original definition has been subject to several revisions in subsequent studies of a similar nature (Vavra, 2000), it is important that a detailed, clear-cut definition of what counts as a T-unit be adopted. Such a definition can be found in Sotillo's (2000) study. The following (slighty abridged and adapted) version of this definition is used for this analysis:

(a) Do not count sentence fragments.

(b) If an NP or subordinate clause is standing alone, do not count them as T-units.

(c) When there is grammatical subject deletion in a coordinate clause, count the entire sentence as one T-unit. Example: *"More and more women take an active part in society and use their ability to help others."*

(d) Count S-nodes with a deleted complementizer as an embedded clause as in: "The main idea of this story is that you can't deny your race or ethnic group and [that] you can't show people how white or how American you are supposed to be."

(e) Count (the German equivalents of) the following as subordinators: after, although, because, if, until, where, since, when, while, as if, as though, so that, in order that, so as, in order, as (many) as, more than, although, even though, despite, so (that).

(f) Count T-units in parentheses as individual T-units.

This definition is rather stringent — criterion (c) in particular considerably restricts the range of sentences that yield a different T-unit length compared to straight sentence length. For this reason, a more lenient version of mean T-unit length has also been used for this analysis. It differs from the strict version only in that criterion (c) is omitted, i.e. sentences with grammatical subject deletion (such as the example sentence in (c)) count as 2 T-units.

Example:
"Die Verbrecher sind gefangen worden und sind jetzt im Gefängnis."
(The criminals have been captured and are now in jail.)

Mean T-unit length (strict): 10
Mean T-unit length (lenient): 5

In this example, the sentence exhibits subject deletion and is thus analysed as consisting of 1 T-unit under the strict definition (and a corresponding mean T-unit length of 10), and 2 T-units under the lenient definition (and a corresponding mean T-unit length of 5).

4.3.3 Coordination Index/ Total Number of Clauses

The Coordination Index (CI), as defined by Warschauer (1996, p. 14), is the ratio of independent clause coordinations, and the total number of combined clauses (independent coordination plus dependent subordination). The CI is considered to be inversely proportional to syntactic complexity, and is essentially a measure of the proportion of subordinate clauses a sentence contains. So for instance, if a sentence contains only a main clause (plus any number of coordinate clauses), the CI is 1.0, whereas a sentence consisting of a main clause plus 3 subordinate clauses would be assigned a CI of 0.25.

The total number of clauses in a sentence is used as a complimentary index. It serves the dual function of providing the basis for the calculation of the CI, and providing a syntactic complexity measure of its own (used for example in Sotillo's (2000) study). However, the measure is relatively crude as e.g. the degree of subordination (embeddedness) is not taken into consideration.

Example:
"Um 610 herum, so viel weiß die Wissenschaft immerhin, formte er den *Teig*, der beim Brotbacken übrig geblieben war, zu einem Gebäck, das aussehen sollte wie betende Kinderhände."
(Around 610 — this much is known to science — he formed the *dough*, which had been left over from baking bread, into pastries which were supposed to look like the praying hands of children.)

Coordination Index for analysis: 2 (Number of Clauses: 4)

In this example, the sentence consists of 4 clauses (1 Simple Main Clause, 1 Parenthesis, and 2 subordinate clauses), with a CI of 2/4 = 0.5. Note that since the CI is inversely proportional to syntactic complexity, the value used for this analysis is 1/CI, i.e. in this example, 1/(2/4) = 4/2 = 2.

4.3.4 "Staircase-Measure"

The extent of subordination that a sentence exhibits is one of the determinants of syntactic complexity; among the indicators introduced so far, it is taken into account by the CI. However, CI only measures subordination by the number, or proportion, of subordinate clauses within a sentence; the *degree* of embeddedness of these subordinate structures is not being considered, even though it is another factor contributing to syntactic complexity. For instance, one of the indicators used by Sotillo is the total number of embedded subordinate clauses; embeddedness is also measured by some of the indicators based on transformational characteristics of a phrase marker, such as the number of embedding transformations in the history of a sentence (Wang, 1970, p. 401).

Since Sotillo's index seems less informative when used for isolated sentences (instead of for longer text passages) due to the relative rarity of embedded subordinate

structures, and with transformational parse analyses not being available, a different sort of index had to be looked for. Since such an index could not be found in the literature, the following measure (dubbed "Staircase-Measure") has been adopted: each level of embedding (a "step on the stair" as it were) has a count value of 1 which is assigned bi-directionally, i.e. in both the 'downstairs' and 'upstairs' direction. This concept is perhaps best illustrated by the following examples:

Consider the three sentences:

(1) The man who traveled on a plane which had been hijacked is seriously ill.
(+1) ↓ _____ ↑ (+1)
(+1) ↓ _____ ↑ (+1)

(2) Here is the man who traveled on a plane which had been hijacked.
(+1) ↓ _____
(+1) ↓ _____

(3) The man who traveled on a plane has a disease which cannot be treated.
(+1) ↓ _____ ↑ (+1) (+1) ↓ _____

All three sentences have the same general makeup in terms of number and type of constituent clauses: one main clause (SFS), and two subordinate clauses (SUB) in each case. Therefore, all of the sentences have the same *Level* and *Coordination Index* values (4 and 1/3, respectively). However, as the examples demonstrate, their level of embeddedness is not the same, which is reflected in different Staircase-Measure values for each sentence: (1) exhibits the highest degree of embeddedness and is assigned a value of 4, (2) has a value of 2, and (3) a value of 3.

Example:
"Um 610 herum, so viel weiß die Wissenschaft immerhin, formte er den *Teig*, der beim Brotbacken übrig geblieben war, zu einem Gebäck, das aussehen sollte wie betende Kinderhände."
(Around 610 — this much is known to science — he formed the *dough*, which had been left over from baking bread, into pastries which were supposed to look like the praying hands of children.)

Staircase-Measure: 5

In this example, the sentence has a Staircase-Measure value of 5, which is derived as follows: +SFS (Part I) –(+1)→ PAR –(+1)→ +SFS (Part II) –(+1)→ SUB1 –(+1)→ +SFS (Part III) –(+1)→ SUB2.

4.3.5 Yngve-Measure

The measure proposed by Yngve (here simply called Yngve-Measure) is based on his "hypothesis of a depth limitation in language" (Yngve, 1960, p. 464). It belongs to the 'constituent-structure' group of measures (i.e. measures that operate on the surface constituent structure of the sentence in question). The measure was proposed by Yngve as part of his model of sentence production where "the depth of embedding of a phrase was the major predictor of processing complexity" (Frazier, 1985, p. 148); however, it has also been interpreted as a model of sentence comprehension "with varying degrees of success in predicting the complexity of understanding and recalling different constructions." (Frazier, p. 149).

Yngve's measure is based on the idea that uttering phrases imposes a burden on speakers' short-term memory, in that they have to keep track of the associated 'commitments' or predictions. Thus, a central prediction of Yngve's model is that left-branching (as well as nesting or self-embedding) structures are more difficult/complex than their right-branching (non-nesting/ non-self-embedding) counterparts, as can be seen in Figure 4.1.

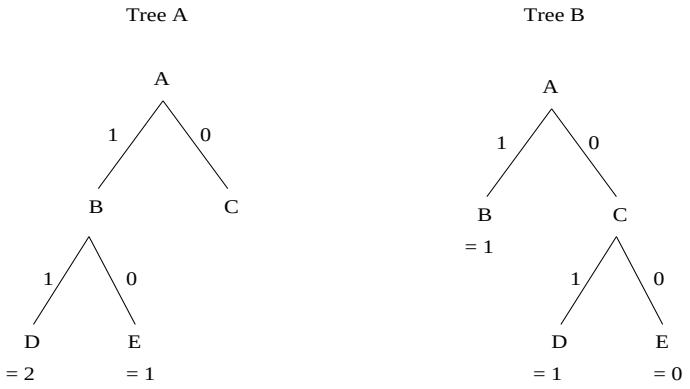


Figure 4.1: Yngve-Measure

Figure 4.1 also indicates how the overall complexity count of a sentence is derived: for any given node, each daughter branch is numbered from right to left starting with zero. The depth of any word (terminal node) is the sum of the numbers of all branches

from the root of the tree to that word; the overall depth of the sentence (taken to be its overall syntactic complexity) is the sum of all word depths. Thus, left-branching tree A in Figure 4.1 has an overall complexity count of 3, while its right-branching counterpart has a complexity count of 2.

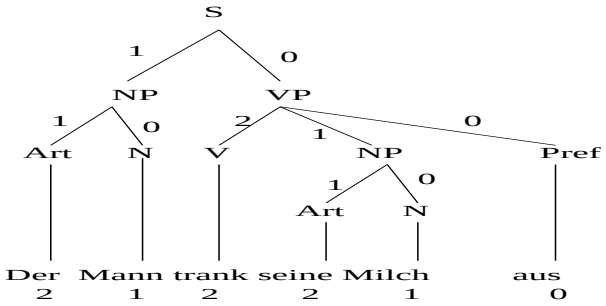
An overview and evaluation of Yngve's measure and its predictions can be found in Frazier (1985, p. 148-155) and Miller and Chomsky (1963, p. 474-75), the most important points of which can be summarised as follows:

(a) Yngve's underlying assumption that the processor knows from the outset the number of daughters that each mother node will dominate is questionable;

(b) Empirical evidence both from language acqusition and adult language processing suggest seems to refute the depth hypothesis's claim that flat, conjoined structures should be perceptually more complex binary branching structures;

(c) Left-branching structures are more complex (difficult to process) than right-branching ones; Frazier notes that while this is partly supported by experimental evidence, a general preference for complex constituents to occur at points of low complexity would suffice to account for this (see (d));

(d) While Yngve's measure correctly predicts that complex constituents optimally occur at points of low depth within a sentence (e.g. the end of the sentence), this insight could be preserved in different models or measures quite distinct from Yngve's. As a case in point, this prediction is also made by most of the measures discussed below.



Example:

"Der Mann trank seine Milch aus."

(The man drank his milk up.)

Yngve-Measure: 8

In this example, the overall complexity count for the sentence is 8. [4]

4.3.6 Non-Terminal-To-Terminal-Node (NTTTN) Measure

This measure, proposed by Miller and Chomsky (1963), arguably provides the conceptually simplest manner of determining the syntactic complexity of sentences among the parse-based measures. In contrast to the Yngve-measure, which only considers the branches of a parse tree, the NTTTN measure only deals with the configuration of the tree nodes. Based on the assumption that syntactic complexity correlates with the amount of superstructure that is associated with the words of a sentence, it simply divides the number of non-terminal nodes in the sentence by the number of terminals, and stipulates the resulting ratio as a measure of syntactic complexity.

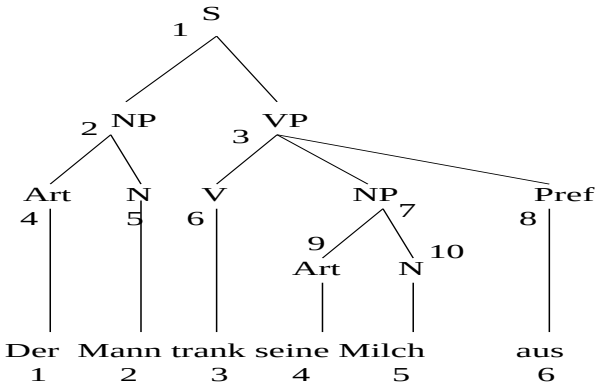
On the one hand, the NTTTN measure avoids some of the problems of the Yngve-measure, as it prefers flat, conjoined structures over binary branching ones, and has no preference with regard to left- or right-branching (cf. Frazier (1985, p. 156)). On the other hand, it fails to make predictions about low *vs* high attachment of a phrase or discontinuous constituents (a shortcoming it shares with the Yngve-Measure that is especially problematic for German, given its ubiquituous syntactic discontinuities). Also, as Frazier (1985) notes, the NTTTN measure is insensitive to the *distribution* of non-terminals over the string which arguably accounts for many differences in processing complexity. An even more fundamental objection is raised by Hawkins (1994, p. 33): because an increase in sentence length (the addition of more terminals) usually entails a corresponding increase in the associated superstructure (the number of non-terminals), the NTTTN ratio can be expected to remain roughly constant across sentences.

Despite these serious drawbacks, the NTTTN ratio has been included in the group of to-be-evaluated complexity metrics mainly on the virtue of its ease of computation.

Example: Non-Terminal-To-Terminal-Measure: 10/6

In this example, the overall complexity count for the sentence is 10/6.

[4] **NB** For this and the following examples, deep structures (with VPs) are used.



4.3.7 Frazier-Measure

As has been noted above, for Frazier the central shortcoming of Miller and Chomsky's NTTTN measure is its non-sensitivity to the precise distribution of non-terminals over the sentence. Her measure of syntactic complexity (Frazier, 1985, p. 157ff.) was an attempt to rectify this problem by introducing a *local* non-terminal count, which also reflects Yngve's observation that complex phrases are easier to process at points of low complexity. Frazier assumes that a local non-terminal count should be computed over a three-terminal window (this assumption appears slightly arbitrary as it seems to be motivated by just one example).

In keeping with the Minimal Attachment Strategy[5], Frazier's measure assumes that non-terminals are introduced only when:

- The first word of a sentence needs to be connected to the matrix S-node;
- Any subsequent word needs to be connected into the current (partial) constituent structure in a way consistent with the phrase structure rules of the language.

Frazier's measure is then "simply the sum of the value of all nonterminals introduced over any three adjacent terminals and thus the maximal local nonterminal count of a sentence is the largest such sum in a sentence" (Frazier, 1985, p. 164).

Hawkins (1994, p. 33-37) raises three main objections to Frazier's measure: (a) there are certain structural types for which the measure makes either no or incorrect predictions; (b) the local focus of Frazier's count is problematic in principle because of the exclusion of a number of non-terminals arguably relevant for determining complexity (i.e. all nodes constructed outside of the three-terminal window); and (c) the

[5]This strategy "specifies that incoming items are attached into a constituent structure representation of the sentence using the fewest nodes consistent with the wellformedness constraints of the language" (Frazier, 1985, p. 135).

equation of *peaks* of complexity with the overall sentence complexity is questionable; Hawkins argues that aggregating over all local complexities would be at least as plausible. However, regarding (c) it should be noted that Frazier does not claim her measure to be necessarily applicable beyond the comparison of sentences with an equal overall non-terminal ratio, and concedes that it should be developed further to allow comparison of sentences with different global nonterminal counts.

Taking these points into account, two versions of Frazier's measure are used for this analysis: Frazier's original version just described (*maximum* over three-terminal windows), and — following Hawkins's suggestion — one where the *average* over all three-terminal windows is computed.

Before the measure is explained with an example sentence, the following points need to be noted:

- On the question of whether all nonterminals contribute equally to the complexity of a sentence, and thus should be assigned the same value, Frazier cites examples suggesting that S and $\bar{S}$ nodes contribute more to sentence complexity than do other node types. Therefore, Frazier assigns these nodes a value of 1.5; all other nonterminals are assigned a value of 1.

- Pre-terminal nodes (e.g. Art, Adv) are excluded from this count, based on the assumption that the syntactic processor does not assign lexical category labels.

.

For the example sentence *Der Mann trank seine Milch aus*, the Frazier measure (max and average) is computed as follows over the string of its 6 terminals:

Der $\rightarrow$ NP (1) + $\rightarrow$ S = 2.5

Mann = 0 (no new non-terminals since N is preterminal)

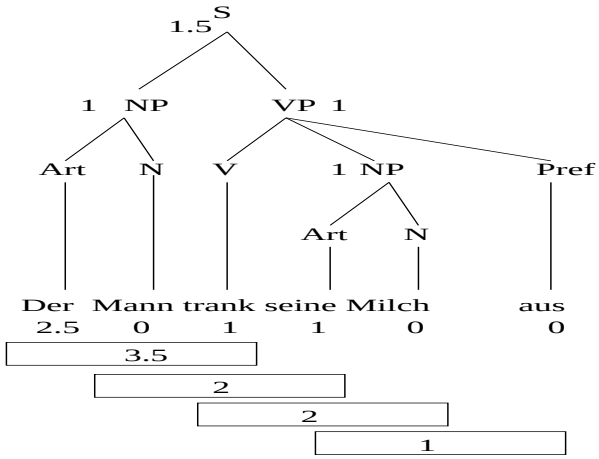
trank $\rightarrow$ VP = 1

seine $\rightarrow$ NP = 1

Milch = 0 (no new non-terminals since N is preterminal)

aus = 0 (no new non-terminals since Pref is preterminal)

As can be seen from above diagram with the 4 three-terminal-windows, the two versions of the Frazier-metric yield the following values for the example sentence:



- **Frazier-Metric (maximum): 3.5**
- **Frazier-Metric (average): 8.5/4 = 2.125**

4.3.8 Early Immediate Constituents (EIC-measure)

Hawkins's (1992; 1994) measure of syntactic complexity is based on the principle of *Early Immediate Constituents* (EIC). The underlying notion of this concept is that "words and constituents occur in the orders they do so that syntactic groupings and their immediate constituents (ICs) can be recognized (and produced) as rapidly and efficiently as possible in language performance" (Hawkins, 1994, p. 57). In other words, Hawkins's central idea is that a phrasal unit is more expensive to process (i.e. syntactically more complex) if it takes longer for a processor to identify its ICs. To illustrate this, Hawkins gives an example of Heavy NP Shift in English:

I vp[introduced np[some friends that John had brought to the party] pp[to Mary]
1 2 3 4 5 6 7 8 9 10 11

I vp[introduced pp[to Mary] np[some friends that John had brought to the party]]
1 2 3 4

It seems obvious that in this example, the first sentence is more difficult to process than the second. In Hawkins's terminology, this is due to the different size of the respective Constituent Recognition Domains (CRDs)[6] of the VP - in the first, 11 words

[6]A CRD is defined by Hawkins as the ordered set of words (relative to a phrasal mother node such as VP, NP) that must be parsed in order to recognize all ICs of that node (Hawkins, 1992, p. 198).

need to be scanned by the processor before all ICs of the VP (namely V, NP and PP) are recognized, whereas 4 words suffice in the second.

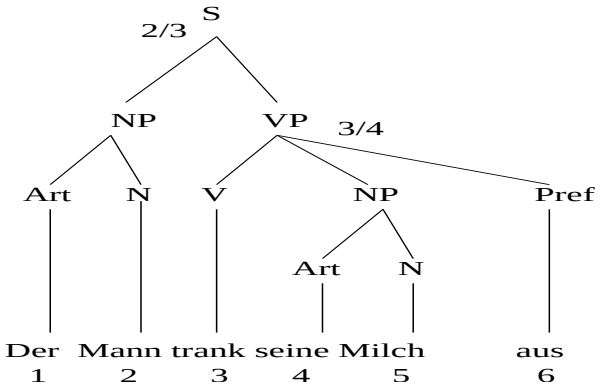
Thus, Hawkins's (1994) EIC-measure is based on the assumption that the human parser prefers to maximize the left-to-right IC-to-word ratios associated with a CRD; the syntactic complexity of any given CRD is then inversely proportional to the IC-to-word ratio of that CRD (which for the example above (VP) would be 3/11 for the first, and 3/4 for the second sentence).[7]

In sum, the EIC-measure shares with Miller and Chomsky's and Frazier's measures the basic assumption that syntactic complexity is a function of the amount of 'superstructure' associated with the terminals of the parse string; it differs from these proposals in the scope of the domains within which complexity should be computed (a CRD as opposed to the whole sentence or three adjacent terminals plus the nodes they construct). By relativizing IC-to-word ratios to CRDs, Hawkins's measure is able to account for the difference in complexity of the two example sentences above (in contrast to Miller and Chomsky's global ratio which fails to do so).

The above raises the question of how the EIC-measure could be extended from CRDs to entire sentences. Hawkins (1992, p. 198) suggests to aggregate over all CRDs in the sentence (i.e. all phrasal categories that the sentence dominates), but unfortunately does not elaborate on this with examples or other demonstrations of the plausibility of said approach. At least for the purposes of this analysis, his proposal seems questionable for the following reasons: (a) it is not evident *a priori* that all CRDs should be assigned the same weight — in fact, it would seem to be more intuitively plausible to assign more weight to VP CRDs than to, say, AdjP ones; (b) on a related point, phrasal nodes that are not Ss or VPs (i.e. NPs, PPs, AdvPs etc) are almost invariably recognised at their first terminal, yielding an IC-to-word-ratio of 100%. To include them in the aggregate count (as Hawkins's suggestion would imply) may therefore unduly distort the overall sentence ratio towards simplicity, especially since their general frequency of occurrence is rather high. Rather than aggregating over all CRDs in the sentence, it seems more plausible to restrict the analysis to those phrasal categories where IC recognition can be reasonably expected to be delayed, namely S, $\bar{S}$ and VP nodes. This is the approach that has been adopted for the analysis task at hand; note that the highest tree node, S, is itself included in the aggregate. In addition to the aggregate version averaging over Ss and VPs, a "peak complexity" version that retains

[7] In an earlier proposal of CRD (Hawkins, 1992), an *aggregate* IC-to-word ratio is suggested which averages over all words of the CRD. For the example above, this would yield 47% for the first and 27% for the second sentence.

only the minimum of these IC-to-word ratios is also included in the analysis.



As can be seen from the diagram above, for the example sentence *Der Mann trank seine Milch aus* there are two relevant CRDs: the S-CRD and the VP-CRD, with IC-to-word ratios of 2/3 and 3/4, respectively. Since the EIC-measure is inversely proportional to syntactic complexity, the values used for the analysis are EIC-measure (peak or average) = 1/(peak or average) EIC-ratio:

- **EIC-Metric (peak): 1/(2/3) = 1.5**
- **EIC-Metric (average): 1/((2/3 + 3/4)/2) = 1.41**

4.3.9 Syntactic Prediction Locality Theory (SPLT)

More recently, the SPLT measure has been proposed by Gibson (1998) as part of his Syntactic Prediction Locality Theory (SPLT), which according to Gibson has been "shown to explain a wide range of processing complexity phenomena not previously accounted for under a single theory [...]" (Gibson, 1998, p. 1). SPLT differs from the measures discussed so far mainly in that it is working-memory based and conceptualizes syntactic complexity as comprising two distinct but related components: a *memory cost* for maintaining or storing syntactic predictions, that is "remembering each category that is required to complete the current input string as a grammatical sentence" (Gibson, 1998, p. 13), and an *integration cost* for integrating new words into the structures built thus far.

Even though both components are part of SPLT, the units of measurement for both are quite distinct: Gibson quantifies the work necessary to perform an integration in terms of Energy Units, where 1 Energy Unit is defined as the product of a Memory Unit and a Time Unit (as indexed by reading times). Since integration cost is reflected

by reading times most directly, which are neither available nor of direct interest for this study, the maximal memory cost alone has been chosen as the SPLT-measure. This choice is also in keeping with Gibson's assumption that "the intuitive complexity of a sentence is determined by the maximal memory complexity reached during the processing of the sentence" (Gibson, 1998, p. 17), and can be justified as a reasonable approximation of the overall SPLT complexity by Gibson's initial assumption that the "memory cost function is the same as the integration cost function: a discourse-based locality function" (Gibson, 1998, p. 14).

This assumption also highlights another important difference of SPLT compared to EIC or the Frazier-measure: SPLT measures distance in terms of the number of new discourse referents (NDRs) that have been processed between the time the syntactic prediction was first made, and the time it is fulfilled. Gibson cites empirical evidence (Gibson and Warren, 1998) for the underlying hypothesis that words introducing NDRs cause a substantially higher increase in integration and memory cost than do other words. In SPLT, a discourse referent "is an entity that has a spatio-temporal location so that it can later be referred to with an anaphoric expression, such as a pronoun for NPs, or tense on a verb for events" (Gibson, 1998, p. 12). NDRs then include NPs with the exception of pronouns as object referents, and main verbs of VPs as event referents. It should be noted that the treatment of NDRs in SPLT is a potential weak point of the measure, or at least an oversimplification of an arguably more complex picture: as Gibson himself points out, "it is also possible that different types of intervening discourse referents cause different increments in memory cost for a predicted syntactic category", and "it may be that not only nouns and verbs cause memory cost increments [...] Adjectives and contentful propositions may also may also cause memory cost increments, because they indicate predications." (Gibson, 1998, p. 25).

Besides NDRs, the second main determinant of the SPLT measure is what exactly should count as a syntactic prediction. In SPLT, no memory cost is assumed for the top-level[8] matrix predicate on the grounds that this prediction is built into the parser (since all well-formed sentences are headed by a predicate, the assumption is that the processor is always expecting a predicate). It is not quite clear, however, why SPLT *does* associate memory costs with the matrix subject, which by the same token should be cost-free as well. Unfortunately it is also not quite clear exactly which syntactic predictions should be included; Gibson's examples seem to indicate that predictions at

[8]This qualification has been introduced to account for clause-based syntactic closure (Gibson, 1998, p. 27-30).

the beginning of phrasal nodes (e.g. article → NP; preposition → PP), are *not* included. Appendix C provides a listing of the syntactic predictions that have been considered for the current analysis, taking the special features of German syntax into account.[9]

The memory cost is provided by the following function[10] which is monotonically increasing and asymptotically approaches a maximum complexity:

$$M(n) = \frac{1}{1 + e^{-n+1}}$$

where M(n) is the memory cost of an item (word) in the parse string relative to a particular syntactic prediction, and n the number of intervening elements (NDRs) for that item and syntactic prediction. The complexity of a sentence is then identified with the maximum value of the total memory costs for each item (summed over all syntactic predictions for each word).

Table 4.1: Example for syntactic predictions of the SPLT measure

Syntactic prediction	**Input word**							
(* indicates point where syntactic prediction is fulfilled)								
	Der	Mann,	der	die	Milch	trank,	ist	dumm.
Matrix verb	0	0	0	0	0	0	0	0
Matrix subject	M(0)	*	0	0	0	0	0	0
RelClause verb	-	-	M(0)	M(0)	M(1)	*	-	-
RelClause subject	-	-	M(0)	M(0)	*	-	-	-
Subject Compl.	-	-	-	-	-	-	M(0)	*
Total Cost	M(0)	0	**2M(0)**	**2M(0)**	M(1)	0	M(0)	0
M(n)≈	0.27	0	0.54	0.54	0.50	0	0.27	0

Table 4.1 illustrates the working of the SPLT measure with the example sentence *Der Mann, der die Milch trank, ist dumm.* (The man who drank the milk is stupid).

The table shows that for this sentence, the memory cost peaks at words 3-5, so 2*M(0) ≈ 0.54 is the SPLT sentence complexity as expressed by its memory cost component.

[9] In general, the principle has been adopted that a syntactic prediction is made at the point where it can be umambiguously predicted 'with the benefit of hindsight'; e.g. verb readings, which may be ambiguous with respect to their prediction at the time of encounter, are disambiguated by the *a posteriori* knowledge of the complete sentence. Even though this is not strictly in keeping with SPLT assumptions, it greatly facilitates the application of the SPLT measure in practice.

[10] This version of the function assumes the default values for its parameters given in (Gibson, 1998, p. 31).

4.4 Empirical Evaluation of the Syntactic Complexity Measures

4.4.1 Introduction

In order to empirically evaluate the measures described in section 4.3, a web-based study has been conducted on a representative sample of the complete set of sentences in the teacher questionnaire data. The purpose of the evaluation study was to answer the following question: which of the measures of syntactic complexity outlined in section 4.3 provides the best correlation with the judgments of (linguistically trained) native speakers of L2 German?

4.4.2 Participants

18 native speakers of German participated in the study. All participants had received at least rudimentary linguistic training at some point during their education and could thus be assumed to be familiar with the basic concepts of grammar. The subjects were recruited over the internet via postings to German and Linguistics departments of British universities, as well as among linguistically trained native speakers of the University of Edinburgh's Division of Informatics and personal acquaintances of the author. All participants were fluent in English and thus had no problem reading the English instructions. Participation in the study was voluntary and unpaid.

4.4.3 Materials

A set of 40 sentences was selected from the complete set of sentence pairs in the teacher questionnaire data (each sentence pair consisting of one original sentence (OS) taken from a newspaper or magazine article, and the corresponding example sentence (ES) provided by the teacher to explain a difficult word in the original sentence — see chapter 3). The 40 sentences were selected in such a way as to strive for a roughly representative sample of the entire range of sentences in terms of syntactic complexity, based on introspection of the experimenter. Despite an attempt to ensure some level of diversity of structures by representing e.g. passive voice and coordination constructions, the selection of sentences was thus largely independent of the actual grammatical structures they contained. Thus, no attempt was made to relate the selection process to the linguistic/psycholinguistic literature on syntactic complexity.

The details of the selection procedure were as follows: for each of the 17 questionnaires, for both the OS and ES section, 3 sentences were picked (based on introspection of the experimenter) as most complex, least complex and average, respectively, yielding 51 sentences for both OS and ES, i.e. a set of 102 sentences in total. Out of this pre-selection set, the final total of 40 sentences to be rated by the subjects was selected in the following way: first, each of the 102 sentences was rated on a scale of 1 to 10 (again based on introspection of the experimenter). Second, for the OS section, 20 sentences were selected evenly across the categories, i.e. 2 sentences were picked for each of the 10 rating categories; for the ES section, the remaining 20 sentences were selected proportionally from the 10 rating categories (see Table 4.2).

Table 4.2: Sentence Selection for Syntactic Complexity Experiment

Rating	**Original Sentences (OS)**		**Example Sentences (ES)**		**Total**	
Category	Rated	Selected	Rated	Selected	Rated	Selected
1	5	2	6	2	11	4
2	4	2	12	5	16	7
3	10	2	12	5	22	7
4	5	2	6	2	11	4
5	3	2	10	4	13	6
6	4	2	2	1	6	3
7	11	2	3	1	14	3
8	5	2	0	0	5	2
9	2	2	0	0	2	2
10	2	2	0	0	2	2
Total	51	20	51	20	102	40

For this phase of the selection process, in the bottom-half rating categories (1-4, indicating low complexity), an effort has been made to ensure diversity in terms of potentially relevant syntactic features (e.g. fair representation of passive voice, verbs

with separated particles, main clause coordinations). The proportional selection in the ES section (and therefore in the final set of 40 sentences) reflects the hypothesised bias (confirmed by introspection) towards low complexity in this section.

4.4.4 Procedure

The study was implemented as an online questionnaire form, consisting of two web-pages. On the front page, subjects were presented with instructions. The instructions (presented in English) first explained the task to the subjects, namely the judgment of the syntactic complexity of 40 sentences on a scale of 1 (least complex) to 10 (most complex). No explanation or definition of the concept of syntactic complexity was offered, but subjects were provided with an example of a syntactically very simple German sentence and an example of a syntactically much more complex German sentence. A rating of 1 was suggested to subjects for the simple sentence, while no rating was suggested for the complex sentence (except for the statement that it was considerably more complex).[11] Subjects were told that there were no 'correct' answers, and that their ratings should *not* be influenced by the semantic content of the sentence (e.g. difficult words or subject matter). Subjects were then asked to adhere to the following procedure: first, to read all items before rating any of them; then, to choose the sentence they thought had the *lowest* syntactic complexity among the items, and to give it a 1; then to choose the most complex sentence and give it a 10; finally, to rate the remaining 38 items on the 10-point scale (possibly, but not necessarily, including the extreme values 1 and 10). This procedure had been chosen so as to ensure that subjects made use of the full range of the scale, including the extreme values 1 and 10. An excerpt of the data section of the questionnaire is provided by figure 4.2.

4.4.5 Results

Prior to further analysis, the inter-rater reliability of the data was assessed using both parametric and non-parametric methods. A very high inter-rater reliability was indicated by the rater *vs* group correlations (group excludes rater), with Pearson's *r* ranging

[11] The rating of 1 for the simple sentence *Markus singt* (Markus sings) could be safely suggested to subjects as no simpler, grammatically well-formed sentence is conceivable than the two-word combination subject plus predicate. The complex sentence chosen was *Niemand weiß genau, was die Katastrophen-Touristen antreibt, doch wo immer sich ein Unglück besichtigen lässt, sind sie, entsprechend ausgerüstet, kurze Zeit später zur Stelle.* [Nobody knows for certain what motivates the disaster tourists, but wherever there is an accident to look at, they will be there — suitably equipped — only a short while later.]

1 Die Verbrecher sind gefangen worden und sind jetzt im Gefängnis.

(The criminals have been captured and are in prison now.)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Very simple ——SYNTACTIC COMPLEXITY——→ Very complex

2 Der Hund bellt, der Tiger schnauft.

(The dog barks, the tiger wheezes.)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

3 Die Firma DHL besitzt 20 Flugzeuge, von denen 10 Frachtmaschinen zum Transport von Waren und Gütern benutzt werden.

(The company DHL owns 20 planes, 10 of which are freight planes used to transport goods.)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Figure 4.2: Translated excerpt from an online syntactic complexity questionnaire page

from 0.81 to 0.96 (average 0.92). This positive result was confirmed by both the Intraclass Correlation Coefficient (ICC), and Kendall's coefficient of concordance. Both individual-rater and group-rater ICCs are very high at 0.82 and 0.99, respectively.[12] Kendall's W uncorrected for ties is W1 = 0.82, while correction for ties yields W2 = 0.85. Both W1 and W2 are significant at $p < 0.01$.

For each of the 40 items, the average and standard deviations σ were computed. The 40x18 data points were then inspected for outliers, which were removed from the data set; a data point was considered an outlier if it was above or below 2 σ's from the item average.[13] From the corrected data set (with outliers removed), new item averages were computed which formed the basis for the correlation analysis (again using Pearson's r).

For each of the metrics under consideration, the correlation with the corrected item averages was computed; the results are given in Table 4.3. The test sentences were divided in a low-complexity and high-complexity group[14], the classification criterion being whether or not the average rating for a given sentence was above or below the composite average rating across all sentences.

4.4.6 Discussion

The inter-rater reliability indices show a high general agreement among the subjects on the construct of syntactic complexity, as well as a high consistency amongst the raters themselves.

The results clearly show that of the MSC analysed, sentence length has the best correlation with average native speaker judgments on syntactic complexity with $r = 0.93$[15].

This result is quite surprising as, of all the MSC tested, sentence length is the most basic and crudest measure, and the only one that does not take any syntactic information into account at all. One might hypothesize that sentence length performs less well among the bottom half of the items (20 least complex sentences), as these sentences tend to vary less in length than sentences in the top half due to the excessive length

[12] The individual-rater version of the ICC is a measure of the typical reliability of a single rater, while the group-rater ICC is an estimate of the correlation of the composite rating of all subjects, and the same type of rating in a re-test. Both versions of the ICC assume that the raters are a random rather than a fixed variable.

[13] 27 out of the 720 data points (ca. 4%) qualified as outliers according to this criterion.

[14] The decision to split the items into just two groups was arbitrary.

[15] This correlation is significant at $p < 0.01$ (one-tailed).

Table 4.3: Correlations of Syntactic Complexity Measures with Average Ratings

(in order of descending *r*)

$^{**}/^{*}$ Correlation is significant at the 0.01/0.05 level (2-tailed)

Syntactic Complexity Measure	*r* **overall**	*r* **top-half (most complex)**	*r* **bottom-half (least complex)**
Number of sentences	**40**	17	23
Average rating	*4.15*	*6.17*	*2.66*
Sentence Length	**0.93****	0.85**	0.92**
Yngve (deep)	**0.90****	0.82**	0.85**
Staircase-Metric	**0.87****	0.76**	0.56**
EIC (peak/deep)	**0.86****	0.77**	0.82**
T-Unit (strict)	**0.83****	0.72**	0.66**
EIC (peak/flat)	**0.83****	0.70**	0.81**
Yngve (flat)	**0.82****	0.69**	0.86**
T-Unit (lenient)	**0.80****	0.80**	0.61**
Coordination Index	**0.80****	0.73**	0.56**
Frazier (peak/flat)	**0.71****	0.40	0.76**
Frazier (peak/deep)	**0.70****	0.51*	0.80**
SPLT	**0.64****	0.37	0.57**
Frazier (average/flat)	**0.45****	0.29	0.35
EIC (average/deep)	**0.37***	0.02	0.19
EIC (average/flat)	**0.35***	-0.15	0.39
Frazier (average/deep)	**0.30**	0.35	0.28
Non-Terminal-To-Terminal (flat)	**0.29**	0.14	0.04
Non-Terminal-To-Terminal (deep)	**0.17**	0.19	-0.09

of some very complex sentences.[16] However, Table 4.3 shows that this is not so: on the contrary, it is the more complex sentences where sentence length performs slightly worse (though still with a very high correlation of $r = 0.85$, and outperforming all other measures in that group). It might be speculated that this is because as sentences get very long (the longest item has 62 words), it becomes more difficult for subjects to adequately 'eyeball' or estimate differences in sentence length.

Also surprising is the overall result that all clause-level measures show very good correlations with native speaker ratings ($r >= 0.80$), performing consistently better than the more fine-grained measures that operate on a constituent-level analysis (it is noteworthy that the only measure among these that also shows a very good correlation — the Yngve-measure — is one of the most basic in that group).

While it is outside the scope of this work to analyse or speculate on the results for the remaining measures in greater detail, the following specific observations can be made:

- All clause-level measures perform significantly better in the top half than the bottom half of the items; this might be explained by the fact that sentences below a certain level of complexity tend to consist of only 1 or 2 clauses, thus not allowing much room for these measures to differentiate;

- The point above is in direct contrast to the performance of constituent-level metrics that show a good correlation with native speaker ratings ($r >= 0.70$), which perform consistently better in the bottom half (for less complex items). This seems to suggest that for less complex sentences, where clause-level analysis does not provide enough differentiating information, subjects tend to base their complexity ratings on constituent structure, whereas for more complex sentences, phrase structures may become too complex for subjects to rely upon, and clause-level analysis alone provides sufficient information;

- The very poor performance of the Non-terminal-to-terminal ratio seems to corroborate Hawkins's objection that this ratio remains largely constant across sentences;

- Of the constituent-level measures that show a good correlation with native speaker ratings ($r >= 0.70$), the difference between the 'flat' and 'deep' versions is

[16] Sentence lengths in the top half (more complex items) show a Standard Deviation of σ = 14.07, whereas the Standard Deviation of sentence lengths in the bottom half (less complex) is only $\sigma = 5.04$.

not significant, although ratings derived from a 'deep' phrase marker perform slightly better overall;

- For both Frazier's measure and the EIC measure, the respective 'peak' versions significantly outperform the 'average' alternatives. It seems that local peaks of complexity are indicative of sentence complexity, and the average versions 'water down' the overall complexity by including too much irrelevant, low-complexity data. However, a different weighting scheme for the average that leans more heavily towards complexity peaks could well lead to considerable improvement.

It should be emphasized at this point that the results of this evaluation of MSC need to be accepted with a certain degree of caution, especially in the context of L2 sentence processing. First, as has been noted in section 4.4.3, sentence selection for the evaluation was based on an introspective pre-evaluation of the sentences of the teacher data and — despite some attempt to ensure some diversity as regards structures such as passive voice, coordination etc — did not sufficiently control for any of the grammatical phenomena and constructions likely to be a factor of syntactic complexity. Second, it cannot be completely ruled out that native speakers that were participating in the evaluation differ in their syntactic complexity judgments from teachers judging the same sentences not necessarily as native speakers, but with the intended target group of their L2 students in mind. Furthermore, as has been indicated above, it remains to be established to what extent native speaker (or even teacher) judgments about syntactic complexity square with the difficulty of sentence processing and comprehension as experienced by L2 learners when reading texts.

These caveats and limitations notwithstanding, the evaluation of different MSC proposed in the literature has shown that the crudest measure of sentence readability, namely sentence length, best correlates with German native speaker judgments of sentence complexity. For this reason, sentence length will serve as the MSC of choice in section 4.5 to assess whether the syntactic complexity of teacher examples is significantly reduced compared to the corresponding original sentences.

4.5 Analysis of Syntactic Complexity for Teacher Data

4.5.1 Significance Analysis

As has been discussed in the preceding section, sentence length has been chosen as the MSC to assess whether a reduction in syntactic complexity of teacher ES (as compared to the corresponding OS) is a significant criterion employed by teachers. In order to investigate this issue, a paired t-test has been conducted on the teacher questionnaire data, which consist of 243 original sentences and their corresponding teacher examples. The t-test reveals that, on average, the sentence length of the example sentences (M=12.44, SE=0.27) is lower than that of the corresponding original sentences (M=20.12, SE=0.64). This decrease in syntactic complexity from OS to ES is highly significant ($t(242)=11.73$, $p<.01$).

4.5.2 Function Fitting Analysis

Having established that the reduction of syntactic complexity is a significant factor in the teacher examples, a function fitting analysis was carried out on the sentence length data in order to predict the suitable length of potential corpus examples from the length of the corresponding sentence in the reading text. The function fitting was performed using least-squares error minimisation (independent variable: OS sentence length, dependent variable: ES sentence length). The functions fitted were linear and polynomial up to the fourth degree; for each of the four fitted functions, a classification analysis was conducted to determine the percentage of ES sentence lengths correctly predicted by the respective function, the criterion of correct classification being whether or not the actual ES sentence length was within the range of predicted sentence length $\pm$ 1 S.D. (=10.21) of the OS-ES length differences. Table 4.4 lists the results of the function fitting after the fit converged (final weighted sum of the squared residuals (WSSR) and the corresponding variance of residuals (WSSR/df)), and the classification analysis (percentage of correctly classified sentence lengths).

As can be seen from the table, the best-fitting linear function (f(x) = -0.939576 * x + 11.1057) already achieves a good fit that is slightly improved upon by higher-degree polynomial best-fitting functions. More importantly, the best-fitting linear function achieves a very high classification rate (97.1%) that is only marginally bettered by the best-fitting quadratic function (97.5%).

Table 4.4: Results of Function Fitting and Classification Analysis

FUNCTION TYPE	Linear	Poly-2	Poly-3	Poly-4
final WSSR	237.0	205.6	192.5	132.8
WSSR/df	6.08	5.41	5.20	3.69
Correct Classification in %	97.1	97.5	97.1	97.1

4.5.3 Discussion

The Function Fitting and Classification analysis has shown that the reduction in syntactic complexity can be satisfactorily modeled by the linear function given above. Given the good fit and high classification accuracy of the teacher data, this function will be used as a pre-filter for the regression model to be developed in chapter 7. This is preferable to using the function as a factor for the model itself, since the model will be developed using logistic regression analysis, which can yield a useful model only if none of its variables perfectly (or almost perfectly) predicts the outcome variable, in this case the helpfulness of the example.[17] However, due to the classification accuracy being close to 100%, this requirement would be violated here.

4.6 Summary

This chapter addressed the question of whether the syntactic complexity of the teacher-provided example sentences has been significantly reduced compared to their corresponding original sentences. To this end, several measures of syntactic complexity suggested in the literature were discussed and empirically evaluated. Of the measures tested, sentence length was found to yield the best correlation with native speaker judgments of syntactic complexity, and was therefore used as the measure of choice in a statistical significance analysis of syntactic complexity reduction. Using a paired t-test, this analysis revealed the syntactic complexity of teacher examples to be significantly reduced compared to the corresponding original sentences. Finally, several function types were fitted to the teacher data to predict the syntactic complexity of the example sentence on the basis of the original sentence's complexity; of these, the best-fit linear function was found to be a satisfactory fit with a very high classification accuracy, motivating its use as a pre-filter for the model of teacher criteria for examples to be

[17] This phenomenon is also known as 'complete separation' (Field, 2005, p. 264).

developed in chapter 7.

Chapter 5

Measuring Sentence Similarity

5.1 Introduction

The analysis of the teachers' explanations in chapter 3 suggested that contextual similarity of the example sentences (ES) as compared to the original reading text was one of the criteria teachers employed in their choice of ES. More specifically, it suggested the concept of sentence similarity[1] between the ES and the corresponding original sentence (OS) as a candidate for further analysis.

The (semantic) similarity between two sentences can be seen as an instance of the general problem of judging the similarity of two pieces of text that may range in size from words to phrases, sentences, paragraphs and entire documents. Lexical similarity, i.e. the issue of word-to-word or concept-to-concept similarity lies at the bottom end of this scale and will be considered in more detail in chapter 6.

Text similarity[2] is a relevant issue for such diverse applications as similarity of documents (e.g. to a given query) in information retrieval, text classification, text coherence[3], text summarization, word sense disambiguation, and machine translation. This diversity of applications raises the question of how text similarity measures ought to be evaluated. As Resnik (1999, p. 95) observes, "the worth of a similarity measure

[1] From here onwards, the concept of *sentence similarity* refers to "semantic" similarity between two sentences (as opposed to e.g. structural similarity), i.e. the degree to which they 'mean' or 'are about' the same things or topics.

[2] Text similarity is an instance of the "similarity problem" in general, that is, the question of how to formalize and quantify the intuitive notion of similarity. This issue has a long history in disciplines such as philosophy, psychology, and artificial intelligence; the plethora of perspectives and similarity measures that have been put forward is staggering and well beyond the scope of this chapter to discuss.

[3] In fact, sentence similarity measures tend to double as measures of text coherence (cf. Foltz et al. (1998); Lapata and Barzilay (2005)) if the latter is understood to refer to local *semantic* text coherence, i.e. leaving out factors such as anaphoric reference etc.

is in its utility for a given task", which in Resnik's view should be "its fidelity to human behavior" (ibid.), in this case human ratings of sentence similarity.

This point of view is also taken in this thesis; human sentence similarity judgments will serve as the gold standard in this chapter against which the measures considered will be judged. However, seeing as, for human raters, "the task of comparing sentence meanings is a difficult one" (Wiemer-Hastings, 2004), the use of this standard has to be justified by satisfactory inter-rater reliability (see section 5.3).

The main purpose of this chapter is twofold: first, to determine a measure of sentence similarity that is suitable for the purpose of this study and empirically validated through correlation tests with sentence similarity ratings of native German speakers; second, to ascertain whether the teachers' ES contain a significantly higher ratio of sentences judged similar to their corresponding OS than what would be found in a random selection.

The remainder of this chapter is structured as follows: section 5.2 provides an overview and discussion of different sentence similarity measures that have been proposed in the literature, with a view to their suitability as candidate measures for this study. Section 5.3 then describes the empirical study on sentence similarity ratings conducted with native German speakers. Sections 5.4 and 5.5 present the two-part analysis of the two measures chosen as candidate measures for the study at hand (Lexical Overlap and Latent Semantic Analysis, respectively): the first part in both sections describes how the measures arrive at their respective scores of the sentences selected for the human raters in section 5.3, while the second part deals with the correlation analysis of the respective scores to the human ratings. Section 5.6 investigates the question of whether the sentence pairs in the teacher data are significantly more similar than randomly selected sentence pairs. Finally, section 5.7 summarizes the chapter.

5.2 Measures of Sentence Similarity

5.2.1 Introduction

Measures of text similarity that have been put forward in the literature can be roughly grouped along two main dimensions: first, the underlying technique and theoretical assumptions associated with the use of that technique; second, the textual scope that

the measure can apply to (either sentence-level[4] or concept-level[5], or both). The latter subgroup, that is, measures that can be applied at both concept and sentence level, tend to be either statistical in nature and produce some measure of text similarity based on two chunks of texts as input; or they are based on concept-level measures that are in some way combined to produce an overall measure of text (sentence) similarity. Since similarity measures at the concept level are discussed in chapter 6, the following overview will focus on measures of text similarity that can be applied as sentence similarity measures, using the first dimension mentioned above as the main classification criterion.

Broadly speaking, sentence similarity measures can be classified into three main groups according to the basis they use to arrive at their sentence similarity measures. The most basic similarity measures are based on some calculation of common elements in sentences A and B; they differ in what exactly counts as a 'common element', and how the common element counts in A and B are combined into an overall sentence similarity score. The basic underlying assumption of sentence similarity measures in this group is that the greater the ratio of common elements to total elements in both sentences, the greater their sentence similarity will be. The second group comprises sentence similarity measures that are based on some taxonomy-based (e.g. *WordNet*) measures of concept similarity. The more sophisticated of these measures are also informed by information-theoretic measures based on corpus counts. The third group can be classified as distributional (or context-based) measures that usually take the form of high dimensional semantic (or vector) space models[6]; crucially, these models are based on the assumption that words with similar meaning tend to occur in similar contexts. These groups of sentence similarity measures are described in more detail below.

[4]Strictly speaking, these measures can be applied to any sequence of words, e.g. paragraphs or entire documents.

[5]For the remainder of the thesis, the term *concept* will be taken to be synonymous to *word sense*, i.e. concept-level measures can be readily transformed into word-level measures after word sense disambiguation has been performed on the text.

[6]An alternative to vector-based distributional measures are probabilistic approaches where similarities between words are expressed via functions over the words' distributional properties. As this description implies, they belong to the group of concept-level measures and are thus discussed in chapter 6.

5.2.2 Measures Based on Common Elements

Sentence similarity measures in this group all share the trait of arriving at their overall similarity score via some calculation of overlap of A and B's common elements; the main difference lies in what counts as a 'common element'.

5.2.2.1 Lexical Overlap

Sentence similarity measures based on some form of lexical overlap are arguably the most straightforward similarity measures available. Being based on some calculation of their common lexical elements — most often some variant of the ratio of shared elements and total number of elements — their conceptual simplicity comes at the expense of being overly simplistic: they all share the characteristic of assigning a similarity score of zero to sentence pairs that have no elements in common, even though sentences with no lexical elements in common may still be semantically related. The most basic measures in this group derive their similarity score from simple word overlap. These measures consider all words in the sentence pair as equally important and would assign a similarity score of zero to such obviously similar sentences as *The physician travelled abroad* vs *The doctor flew to Spain* due to their ignorance of synonyms.

Enhanced versions of lexical overlap measures attempt to rectify this shortcoming by either expanding the range of words that count as common elements (e.g. by extending the concept of matching words to the word's synonyms, or to even more distant lexical relations such as hyponyms and antonyms), or by excluding stopwords (common or function words) that can be assumed to contribute little to the overall similarity. More sophisticated improvements on basic lexical overlap add various weighting and normalization factors to the similarity formula, or they may include some brevity penalty to penalize short-length sentences (e.g. Papineni et al. (2002)).

While these improvements on basic lexical overlap rectify the principled shortcomings of these measures to some degree, they still fail to account for local word order as well as "deeper-seated" structure-dependent factors of sentence similarity; however, previous research has shown that relational similarity also has an effect on human ratings of structured scenes (Goldstone, 1994) and sentence similarity (Wiemer-Hastings, 2004), and that "when determining the similarity of texts, human raters apparently tend to ignore similarities between segments with different functional roles" (Wiemer-Hastings, 2004).

5.2.2.2 Common *N*-gram methods

Another extension to basic lexical overlap measures are precision-based overlap measures considering common *n*-grams. By combining common *n*-gram scores for *n*-grams of different lengths (e.g. unigram to bigram in (Shimohata, 2004); unigram to 4-gram in (Papineni et al., 2002)), they are able to take local word order into account (while still failing to account for deeper structural relations). These methods are commonly used for machine translation tasks (Papineni et al., 2002) and require the existence of a reasonably-sized corpus of reference translations (rather than a single reference translation sentence), in order to cancel out effects of different wordings of the same concept (e.g. *East African economy* vs *economy of East Africa*).

5.2.2.3 Suitability of Measures Based on Common Elements for this study

Despite their obvious shortcomings, measures based on lexical overlap are attractive candidates as measures for this study because of their conceptual simplicity and (relative) computational inexpensiveness. Studies such as (Wiemer-Hastings, 1999) have shown that simple keyword matching performs surprisingly well in terms of their correlation with human judgments of sentence similarity ($r = 0.40$), with enhanced versions almost approaching the performance of complex statistical models such as LSA (see below). They will be used as candidate measures in section 5.4 in both their most basic and enhanced versions that address the above mentioned drawbacks at least to some extent. Common *n*-gram methods, however, will not be considered as candidates because of their increased computational costs and, more importantly, their dependence on a *corpus* of reference sentences rather than the single original sentence available for the data at hand.

5.2.3 Taxonomy-based Methods

Most taxonomy-based measures derive the sentence similarity score in two steps: first, they employ a lexical-semantic network such as *WordNet* to arrive at similarity measures at the concept (word sense) level; second, they combine the sense-level similarity scores of the sentence pair's words into an overall sentence-level similarity score.

Concept-level measures mainly differ in whether and to what extent they enhance the taxonomy-derived information by information-theoretic measures based on statistical corpus analyses (see Budanitsky and Hirst (2001) for an overview).

The calculation of the overall similarity score of the sentence pair then typically takes the form of some aggregate of the pairwise concept-level similarities of the two sentences, and the product of their respective number of words, as in Lapata and Barzilay's (2005) formula

$$sim(S_1, S_2) = \frac{\sum\limits_{\substack{w_1 \in S_1 \\ w_2 \in S_2}} \operatorname{argmax}_{\substack{c_1 \in senses(w_1) \\ c_2 \in senses(w_2)}} sim(c_1, c_2)}{|S_1||S_2|}$$

A slightly different tack has been taken by Li et al. (2004), who also derive their sentence similarity algorithm first from a word-level semantic similarity score based on a combination of taxonomic and corpus-statistical information, but then enhance it by a measure of 'word order similarity' based on the position of word appearance in the sentence, which is given a lower weight than the semantic similar score. However, it is not clear how this method compares to human ratings (or other sentence similarity measures, for that matter), as the authors have only evaluated its usefulness (on a very restricted data set) in terms of its ability to separate wanted from unwanted sentences in the domain of conversational agents.

Taxonomy-based measures are not considered as sentence similarity measures in the remainder of this chapter, since (a) they rely on taxonomic information from a lexical-semantic network whose coverage for German is not sufficient at present (Wordnet's German equivalent, GermaNet, currently covers only a fraction of WordNet's taxonomic information); and (b) their use would be too computationally expensive for the purposes of this study, requiring as they do the combination of taxonomic and information-theoretic, corpus-based information[7] into an overall measure of sentence similarity.

5.2.4 Vector-Space Measures

Most recent work in both Information Retrieval and Computational Linguistics has approached the issue of text or concept similarity from a distributional perspective, typically in the guise of semantic (or vector) space models of word co-occurrence. These models collect statistics about the relative frequency with which words appear "near" other words (where the scope of "near" may vary from model to model), and

[7] Taxonomy-based approaches enhanced with information-theoretic measures have been shown to perform best as measures of semantic relatedness, as evaluated in tasks such as malapropism detection (Budanitsky and Hirst, 2001).

represent words (or, in some models, sets of words) as vectors situated in a high-dimensional semantic space. Similarity between words can then be expressed by some suitable vector similarity measure (usually the cosine of the two vectors, see below).

The central idea underlying these models may be pithily summarized by Firth's (1957) well-known quote "You shall know a word by the company it keeps". It needs to be emphasized, however, that distributional similarity measures go beyond mere co-occurrence analysis. Rather, they are based on the notion that words are similar if they occur within similar contexts, that is, they tend to occur in passages which are 'on the same topic'. This has led Higgins (2005) to characterize vector-space methods as based on the 'topicality assumption' (see also the discussion in chapter 6). Landauer et al. (referring to Latent Semantic Analysis) have expressed this idea as "[...] the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of words and sets of words to each other" (Landauer et al., 1998). This approach makes it possible for vector space models to treat words (or sets of words such as sentences, paragraphs etc) as similar to each other even if they do not share any lexical items.

A characteristic common to the three most widely cited vector-space methods — Latent Semantic Analysis (LSA)[8], HAL and Random Indexing — is that they treat and represent context as a "bag of words", i.e. they derive their vector representations from a representation of text based on frequency co-occurrence statistics about which words appear near other words. As a consequence, these models are agnostic about linguistic information such as word order, morphology or syntactic relations[9], a shortcoming that is also the most frequently leveled criticism against them. For example, the disregard of word order would lead these models to assign the maximum similarity score to sentence pairs such as [*The quick brown dog jumped over the fox; The quick brown fox jumped over the dog*].

Recently attempts have been made to address this problem by including syntactic and relational information that the traditional vector-space models have left by the wayside, either by incorporating it within the LSA framework (see below), or by constructing a different type of semantic or vector space that is not based on "bag-of-words"-style co-occurrence counts. One such attempt was proposed by Padó who used syntactically parsed data to construct a context reflecting dependency grammar rela-

[8] also known as Latent Semantic Indexing (LSI) when used for Information Retrieval tasks

[9] This is not strictly true for HAL, whose row and column vectors represent the right and left context of a word; however, this information is discarded in later stages of the algorithm.

tions between words (Padó, 2002; Padó and Lapata, 2003).

In the following, the discussion will center on LSA as it is the most widely cited and used of the vector-space methods, in particular with respect to human judgments of text similarity. The other methods of note, HAL (Lund and Burgess, 1996) and Random Indexing (Sahlgren, 2001), will be contrasted to LSA where appropriate.

5.2.4.1 Latent Semantic Analysis (LSA)

LSA can be conceptualized as either a theoretical model of the representation and computational processes underlying human acquisition and use of language knowledge (e.g. Landauer and Dumais (1997)), or as a practical method of approximately measuring the similarity between two pieces of text. It is this second perspective, in particular in relation to LSA's usefulness as a measure of sentence similarity, that is of interest in the context of this chapter; no position is taken on its merits as a cognitive model. Since the underlying computational mechanisms of LSA have been described in detail elsewhere (Deerwester et al., 1990; Landauer et al., 1998; Wiemer-Hastings, 2004), the following description of LSA will focus on a brief summary of its workings.

The first stage of LSA's mechanism consists of a two-step process operating on a large sample of machine-readable text (usually on the order of a book) separated into "documents"[10], and then representing this training corpus as a word-to-document matrix where each row stands for a unique word (form), and each column for a text passage or document. The cells contain the frequency counts of how many times the respective term[11] appears in the passage denoted by its column. Then, a pre-processing step is applied where the cell frequencies are weighted by an information-theoretic function (typically "log entropy") that reduces the bias of common words by taking into account their respective information gain.[12]

The next step applies the (computationally intensive) matrix algebra technique of singular value decomposition (SVD) to the co-occurrence matrix, which has the effect of reducing the number of dimensions by retaining only the most significant ones. The purpose of this reduction step is to arrive at a representation which is supposed to

[10] "For most applications, each paragraph is treated as a separate document based on the intuition that the information within a paragraph tends to be coherent and related." (Wiemer-Hastings, 2004)

[11] In LSA parlance, a term is a word that occurs in at least two documents; words that occur only once are not represented in the matrix.

[12] HAL differs from LSA at this stage in that it is a document-space rather than a word-space model, that is, it constructs a word-to-word matrix (i.e. there is no need to pre-segment the training text into documents), using a sliding context window that is passed over the corpus to calculate the co-occurrence counts.

capture assumed "latent" dimensions of word meanings that are hidden or obscured in the first-order co-occurrence representation. In terms of the vector representation, this dimensonality reduction enables words that tend to occur in similar contexts to have similar vectors and therefore achieve a high similarity score.[13] The question of which number of dimensions to choose has to be settled empirically — typically around 300 dimensions (ca. ± 100 depending on the corpus and application domain) have been shown to capture the meanings of texts well.

Having arrived at the reduced representation in the high-dimensional vector space, the similarity of text pairs (words or passages) can then be computed by comparing their respective vectors, usually employing the cosine metric for this purpose as it has been shown to work well empirically (see Rehder et al. (1998) for a comparative analysis of vector cosine with alternative measures).

LSA has been shown to have a high correlation with human behavior and textual similarity judgments, such as synonym identification in a multiple-choice setting such as TOEFL's synonym test section (Landauer and Dumais, 1997), essay grading (Landauer and Dumais, 1997), measuring the textual coherence of student essays (Foltz et al., 1998), or comparing a student's answer for a question to a set of expected answers (Wiemer-Hastings, 1999). Despite the generally impressive performance in these tasks, LSA seems to achieve the best results for either word or passage similarity judgments (i.e. either single words or longer texts), with steadily increasing performance for more than 60 words (Wiemer-Hastings, 1999). Rehder et al. (1998) have shown that for 200-word essay passages, LSA accounted for 60% of the variance in human scores, while for 60-word essay segments this figure dropped to only 10%, and even less than 10% for sentence-length segments. Wiemer-Hastings (1999) has reported a maximum correlation of LSA ratings of student responses in an Intelligent Tutoring System of $r = 0.48$ to human ratings, which is only marginally better compared to ca. $r = 0.40$ for the simple keyword method.

This apparent weaker performance of LSA in the task of sentence similarity judgments has been attributed to the "bag-of-words"-approach that takes neither taxonomic or syntactic relations (not even word order) into account. A related shortcoming concerns the filtering of stopwords, e.g. frequent function words such as *not*. This is undesirable for the treatment of negation, as the sentence pair [*John did not hit Mary; John did hit Mary*] would receive a maximum similarity score by LSA. Relating back

[13] Random Indexing does not require the reduction step of SVD, as it uses index and label vectors for each word instead of constructing a huge co-occurrence matrix.

to the criticism of LSA's lack of syntax, Li et al. (2004) note that the use of stopwords is problematic for measures of sentence similarity, since they tend to carry syntactic information that cannot be ignored at the level of short texts such as sentences.

Wiemer-Hastings and Zipitria (2001) suggested that the poorer performance on the sentence level may be due to longer texts providing enough "context cover" that considerably lessens the detrimental impact of the lack of syntactic details. This has led to several attempts to "enhance" basic LSA (for English) with relational information (Wiemer-Hastings, 2004), or syntactic information, e.g. by adding parts-of-speech tags or segmenting sentences in their basic grammatical roles such as subject, verb and object (Wiemer-Hastings, 2000; Wiemer-Hastings and Zipitria, 2001; Kanejiya et al., 2003; Serafin and Di Eugenio, 2004). However, these attempts have so far failed to produce significantly better correlations to human judgments compared to basic LSA versions; Wiemer-Hastings and Zipitria (2001) have reported a correlation to human raters which was only slighty better than that of standard LSA ($r \simeq 0.55$). Overall, the empirical validity of this approach remains to be demonstrated.

Other notable criticisms of LSA relate both to practical inadequacies, and to its status as a theoretical model of language acquisition and utilization. The most notable of these concerns the fact that despite LSA's generally impressive performance on human text similarity judgment tasks, it is still the case that "many LSA word-to-word and passage-to-passage relations will not correspond to human intuition" (Landauer, 2002). Apart from the already noted "bag-of-words"-related deficiencies, this tends to be attributed to the use of non-optimal corpora, which are necessarily "always smaller and different from the total language exposure of any one person" (ibid.). Currently, computational limitations prevent the use of huge corpora due to the complexity of the SVD algorithm.

Another criticism of LSA pertains to the "black-box" character of LSA's mechanism, which may render LSA suspect as a theoretical model as outlined above. This concerns the fact that the dimensions of the resulting model after compression have to be chosen empirically, and are not interpretable in terms of any 'real' concepts or features.

Another open question regarding the training corpus for LSA concerns the benefits of using a lemmatized version of the training corpus, as LSA in its basic guise does not use stemming or other morphological analyses. However, combining different forms of the same lemma can be hypothesized to have a beneficial effect on LSA's performance, since lemmatizing the training text increases the amount of conceptual

information available to LSA. The benefit of a lemmatized corpus is arguably also dependent on pertinent characteristics of the language at hand, such as its position on the spectrum from synthetic or agglutinative to non-synthetic (non-agglutinative) languages.[14] Zipitria et al. (2006) have investigated the effect of lemmatization for both agglutinative (Basque) and non-agglutinative languages (Spanish) and found that using a lemmatized corpus seems to yield greater improvements for Basque, while Spanish seems to function better without lemmatization.

The size and form of the training corpus have been shown to have a non-negligible influence on the success of LSA as a measure of text similarity. As regards corpus size, studies such as (Wiemer-Hastings et al., 1999) and (Olde et al., 2002) have shown that while, in general, an increase of the size of the relevant training text is beneficial, corpus size only has a modest impact on LSA performance, and clearly there seems to be no linear relation between corpus size and LSA performance. Wiemer-Hastings et al. (1999) found no significant difference between the 1/3 and 2/3 versions of the full training corpus (2.3 MB), while Olde et al. (2002) note that "a relatively small amount of relevant material can produce acceptable performance with LSA" (Olde et al., 2002, p. 711).

With respect to the *form* of the corpus, LSA researchers concur on the importance of the *naturalness* and *relevance* of the corpus data as general guidelines for the selection of the training corpus (Landauer, 2002; Olde et al., 2002).[15] On the other hand, addressing the potential benefit of eliminating irrevelant 'noise' data in the corpus (e.g. irrelevant material, event listings in newspaper corpora etc), Olde et al. (2002) found that there was no significant payoff in sanitizing the corpus.

5.2.4.2 Suitability of Vector-based Methods for this study

Despite its shortcomings discussed above, LSA has been chosen as a candidate measure for this study. It is the most widely used and cited of all statistical vector-spaced measures, and has been used extensively (and with good results) on all sorts of text similarity tasks. It is also attractive from a practical point of view, since it is language-independent and straightforward to train and use: besides a machine readable training

[14] Agglutinative languages (e.g. Basque) are a form of *synthetic* languages (languages with a high morpheme-to-word ratio such as German). In agglutinative languages, words are formed by joining morphemes together, resulting in a great amount of word variability. This means that lemmatization reduces the number of LSA terms much more drastically in agglutinative languages such as Basque compared to non-agglutinative ones like Spanish.

[15] For example, it would be ill-advised to use an ecomomy text as training material for LSA with a view to judging student essays on physics.

corpus, no other resources, such as a parser or semantic network, are required. Even though the SVD compression step is admittedly computationally quite expensive, the training cost is a one-time cost, justifying the consideration of LSA as a measure for relatively constrained real-world applications such as Vocabulary Learning in an ICALL environment. These traits make (standard) LSA more attractive as a candidate measure than the alternative models enhanced with syntactic or relational information discussed above, in particular since, for the latter, the empirical validation in terms of significantly improved correlations to human ratings is still outstanding.

Besides its practical expedience, it is mainly LSA's potential ability to achieve high correlations with human similarity judgments that makes it attractive as a candidate measure for this study. Even though LSA's performance on text similarity tasks is not as strong for sentences as it is for words and longer texts, it has to be pointed out that (a) these results were achieved with a smaller corpus than the one available for this analysis (see below), and (b) that the investigated language in question was English. It may well be speculated that LSA can achieve better results for German than it did for English, since German is a synthetic, i.e. morphologically "richer", language than English. Thus there is reason to believe that the syntactic "who did what to whom?"-type information not available for English may be, at least to some limited extent, available for German. Also, the fact that standard LSA fails to exploit word order information may be less detrimental for German, since word order in German is not as tightly constricted as it is for English. However, this potential gain in syntactic information that is provided by a synthetic or agglutinative language may be counterbalanced by the loss of conceptual information caused by the increased 'fragmentation' of a lemma into its inflected forms.

5.3 Empirical Study: Sentence Similarity Judgments of Human Raters

5.3.1 Introduction

In order to empirically evaluate the sentence similarity measures chosen for analysis in the preceding section (lexical overlap and LSA), a second web-based study has been conducted on a representative sample of the complete set of sentence pairs in the teacher questionnaire data. The purpose of this study was to derive a set of human sentence similarity ratings whose average would then serve as the gold standard against

which the similarity measures could be evaluated.

The remainder of this section describes the design and results of this study, in particular with regards to the inter-rater correlations achieved by the participants.

5.3.2 Participants

18 native speakers of German participated in the study. Of these, 9 participants were male adults or near-adults, and 9 were female adults. The subjects were recruited over the Internet via postings to German departments of British and German universities, as well as among family relations, personal acquaintances of the author and German native speakers in the University of Edinburgh's Division of Informatics. All participants were fluent in English and thus had no problem reading the English instructions. Participation in the study was voluntary and unpaid.

5.3.3 Materials

A set of 40 sentence pairs was selected from the complete set of sentence pairs in the teacher questionnaire data (each sentence pair consisting of one original sentence (OS) taken from the reading materials used for the study described in chapter 3, and the corresponding example sentence (ES) provided by the teacher). The 40 sentence pairs were selected from the complete set of teacher sentence pairs *prior to* the exclusion of multi-word items and definitions from the set (see chapter 3). This meant that 9 out of the 40 selected sentence pairs were not part of the final set of teacher sentence pairs due to their containing multi-word target words. Since the OS were extracts from a larger text, they may contain anaphoric references (mostly in the form of pronouns) to preceding sections of the reading text. In order to ensure maximum comprehensibility of the presented sentence pairs for the participants, for all selected OS, any such references (for 12 out of the 40 selected sentence pairs) were replaced with their corresponding contextual referents as they appeared in the preceding texts.

The 40 sentence pairs were selected in such a way as to ensure they constitute a representative sample of the entire range of sentence pairs in terms of their similarity, the selection procedure being as follows:

For each of the 17 questionnaires, 6 sentence pairs were selected in the following way: two sentence pairs were picked (based on introspection of the experimenter) as most and least similar, respectively; the remaining 4 sentence pairs were chosen so as to roughly represent the intermediate sections of approximately average complexity. Out

of this pre-selection set of 102 sentence pairs, the final selection of 40 sentence pairs was arrived at by rating each of the 102 items on a Likert-type scale of 1 to 10 (again based on introspection of the experimenter). Then, 4 items from each of the 10 rating categories were selected in such a way as to represent each of the 17 questionnaires with either 2 or 3 sentence pairs. A complete listing of the 40 selected sentence pairs (after anaphora resolution) is provided in Appendix D.[16]

The experiment was implemented as an online questionnaire form, consisting of the instruction page and (appearing as a separate window) two pages containing 20 sentence pairs each. The decision to split up the data items into 2 groups was due to the assumed difficulty for participants to compare and rate 40 sentence pairs in one go. The items were doubly randomized in the following way: for every sixth participant, all 40 sentence pairs were randomly allocated into 4 groups A-D, such that each group contained exactly one item for each of the 10 rating categories. For the first participant in each six-participant group, groups A and B comprised the first set of 20 items, groups C and D the second. The third and fifth participants were presented with two different group permutations (keeping group A constant in the first set), while the second, fourth and sixth participants were presented with the two sets of their respective predecessors in inverse order. In sum, group allocation for every six-participant group was (AB-CD, CD-AB, AC-CB, CB-AC, AD-BC, BC-AD). For each participant, the items in each 20-item set were presented in random order.

The instruction page of the questionnaire is presented in Appendix E, the second page with a 20-item set to be rated can be found in Appendix F. A translated excerpt of the data section of the questionnaire is provided in figure 5.1.

5.3.4 Procedure

The instructions (presented in English) first explained the task of judging the semantic similarity of 2x20 sentence pairs on a Likert-type scale of 1 (least similar) to 10 (most similar). No definition of the concept of semantic similarity of two sentences was offered, except that subjects were told that semantic similarity referred to the extent to which two sentences "mean" or "are about" the same thing(s), as opposed to their

[16] One of the sentences presented to subjects contained a typo in the form of a word appearing twice, which has been corrected in the appendix. The error occurred in the sentence *Seine Stimmung schwankte immer zwischen* **immer** *den 2 Extremen Euphorie und Schwermut* (His mood always oscillated between **always** the two extremes euphoria and gloom). As this is an obvious typo with no apparent bearing on the meaning of the sentence, it appears safe to assume that it did not affect the similarity ratings of the participants.

1 Bei manchen kommt das Gefühl der Angst immer wieder - regelmäßig und zerstörerisch.
(Fear keeps coming back to some of them - regularly and destructively.)
Fußballfans sind oft zerstörerisch, nachdem sie verloren haben -
sie machen dann Dinge kaputt.
(Football fans often act destructively after they have lost a match - they vandalize things.)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

lowest similarity ——-SENTENCE SIMILARITY——-→ highest similarity

2 Ich glaube, es fängt schon mit einem bestimmten Grundton an, dem Ton der Häme.
(I think it already starts with a certain basic tone, the tone of malicious joy.)
Er war schadenfroh wie immer und betrachtete ihr gebrochenes Bein voller Häme.
(He was gloating as usual and looked at her broken leg with malicious joy.)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Figure 5.1: Translated excerpt from an online sentence similarity questionnaire page

syntactic or structural similarity. Subjects were also told that every sentence pair would have at least one word in common (the target word), and that they should disregard this fact and assign a '1' to the item they thought had the lowest similarity among the 20 items presented on each page. The participants were also provided with two examples of a very similar and a very dissimilar sentence pair (the selections were based on the judgment of the experimenter — see Appendix E). They were told the examples should receive a relatively high and relatively low rating, respectively (i.e. no specific ratings were suggested for the examples).

The subjects were then asked to adhere to the following procedure: first to read all 20 items (in the first data set) before rating any of them; then, to choose the item they thought was the least similar, and give it a 1; then, to choose the most similar sentence pair, and give it a 10; and finally, to rate the remaining 18 items on the 10-point scale (possibly, but not necessarily, including the extreme values 1 and 10). This procedure had been selected so as to ensure that subjects made use of the full range of the scale. The participants were also told that after completing the first data set of 20 items, they had the option to either proceed to the second set, or to exit and submit their ratings at that stage.

For each 20-item set, the ratings could be changed until subjects submitted their ratings, or proceeded to the second 20-item set. No time limit was set for either the item presentation or for the response (subjects were told the survey would take approximately 20-30 minutes). The time subjects took for completing the questionnaire was not recorded.

5.3.5 Results

As part of a pre-analysis of the data obtained, the data of one participant were eliminated after an inspection of her responses indicated a bimodal rating distribution, and a preliminary statistical analysis of this subject's data provided further confirmation that the subject had not completed the task adequately.[17]

Of the 17 participants whose data were retained for further analysis, 14 subjects submitted ratings for the complete set of 2x20 items, while 3 subjects submitted ratings for the first set only.

Prior to further analysis, the inter-rater reliability of the data was assessed using both parametric and non-parametric methods. A moderately high inter-rater reliability was indicated by the rater *vs* group correlations (group excludes rater), with Pearson's r ranging from 0.56 to 0.89 (average $r = 0.8$). All of the rater-*vs*-group correlations were significant at $p < 0.01$. This result was confirmed by the Intra-Class Correlation Coefficient (ICC), which takes differences due to raters into account and estimates the average correlation among all possible orderings of pairs.[18] The two-way random ICC[19] was computed for both single measures (which estimates the reliability of a single rater), and average measures (which estimates the correlation the composite rating of a group of raters, and the same type of rating in a re-test), using both an absolute agreement and a consistency definition.[20] All four versions of the ICC are reported in table 5.1.

Kendall's coefficient of concordance W[21] uncorrected for ties was $W1 = 0.63$, while correction for ties yielded $W2 = 0.67$; both W1 and W2 were significant at

[17] The statistical analysis showed a highly significant negative correlation with the group rating average ($r = -0.89, p < 0.01$), and a skewed (high) average of 6.95 (compared to the overall average of 5.5) indicating rater bias.

[18] The ICC ranges from 0 to 1 with 1 indicating perfect reliability.

[19] Both item and rater effects are considered random

[20] In the consistency definition, the between-measure variance is excluded from the denominator variance.

[21] Kendall's W ranges from 0 to 1 and can be interpreted as a coefficient of agreement among raters. As it is based on ranked data and thus requires the same number of data points for each subject, only data from the 14 subjects that had submitted ratings for all 40 items were considered.

Table 5.1: Intraclass Correlation Coefficients

ICC	absolute agreement definition	consistency definition
single measures	0.745	0.775
average measures	0.990	0.992

$p < 0.01$.

The data points were then inspected for outliers, which were removed from the data set; a data point was considered an outlier if it was beyond $\pm$ 2 standard deviations from the item average. From the corrected data set, the item averages were computed which formed the basis for the correlation analyses for the lexical overlap and LSA methods described below. Table 5.2 provides the mean ratings for each of the 40 sentence pairs in the test set (the sentence pair numbers correspond to those given in Appendix D).

5.3.6 Discussion

The high levels of average rater-group correlation ($r = 0.80$) and two-way $ICC = 0.78$ (two-way random, single measures using consistency definition) found in the participant ratings are on a similar level as the best inter-rater correlations reported in similar studies of human ratings of sentence similarity (Wiemer-Hastings (1999); Zipitria et al. (2006)). The analyses indicate that there is general agreement among the raters on the construct of sentence similarity, and that the raters are mostly consistent across themselves. This result validates the use of the average sentence pair ratings as the gold standard against which the measures of lexical overlap and LSA are compared in the following sections.

5.4 Analysis of Sentence Similarity Teacher Data with Lexical Overlap

The following versions of the lexical overlap measure have been analyzed in terms of correlations with the human ratings of the 40 test sentence pairs:

Table 5.2: Sentence pair mean ratings (in descending order)

Sentence Pair	Mean Rating	S.E.	Sentence Pair	Mean Rating	S.E.
Pair # 34 (N=14)	9.93	0.07	Pair # 37 (N=14)	3.21	0.62
Pair # 21 (N=14)	8.93	0.47	Pair # 9 (N=14)	3.07	0.36
Pair # 4 (N=15)	8.73	0.27	Pair # 16 (N=15)	2.60	0.52
Pair # 18 (N=13)	8.54	0.40	Pair # 15 (N=14)	2.57	0.37
Pair # 36 (N=16)	7.75	0.47	Pair # 32 (N=16)	2.44	0.24
Pair # 1 (N=17)	7.18	0.69	Pair # 11 (N=15)	1.93	0.28
Pair # 28 (N=14)	7.07	0.55	Pair # 29 (N=15)	1.93	0.21
Pair # 8 (N=15)	6.60	0.77	Pair # 27 (N=16)	1.81	0.23
Pair # 25 (N=15)	6.53	0.76	Pair # 23 (N=15)	1.80	0.22
Pair # 19 (N=14)	6.29	0.77	Pair # 26 (N=14)	1.79	0.21
Pair # 30 (N=15)	5.93	0.52	Pair # 10 (N=16)	1.69	0.20
Pair # 7 (N=14)	4.86	0.71	Pair # 5 (N=15)	1.60	0.19
Pair # 20 (N=16)	4.81	0.47	Pair # 24 (N=15)	1.60	0.19
Pair # 2 (N=17)	4.76	0.55	Pair # 31 (N=15)	1.60	0.13
Pair # 39 (N=15)	4.40	0.40	Pair # 33 (N=15)	1.47	0.13
Pair # 17 (N=16)	4.31	0.62	Pair # 22 (N=17)	1.41	0.12
Pair # 12 (N=13)	4.23	0.34	Pair # 3 (N=13)	1.38	0.14
Pair # 35 (N=14)	4.21	0.64	Pair # 38 (N=14)	1.21	0.11
Pair # 6 (N=15)	3.60	0.52	Pair # 40 (N=15)	1.20	0.11
Pair # 14 (N=16)	3.31	0.48	Pair # 13 (N=14)	1.00	0.00

1. *Basic lexical overlap*. This version is the most basic implementation of the lexical overlap method, as it is based on a simple count of common word types. Excluded from the word count are only articles and the basic auxiliary verbs *sein* (to be) and *haben* (to have).

2. *Content words only*. This version excludes frequent function words; in addition to the above stopwords, this includes pronouns, prepositions, and numerals from the common word count.

3. *Synonym-enhanced basic overlap*. This version is based on version (1) but enhances it with the inclusion of synonyms in the common word count. For the purposes of this analysis, two words are considered synonyms if they are listed as such in one of the following lexicographic resources: GermaNet (the German version of WordNet), *Wortschatz Universität Leipzig*[22], and the monolingual dictionaries *WAHRIG*, *DUDEN Bedeutungswörterbuch*, *PONS Großwörterbuch*, and *Langenscheidt Großwörterbuch Deutsch als Fremdsprache*.

4. *Synonym-enhanced content words-only overlap*. This version is a combination of approaches (2) and (3).

For the purposes of this analysis, multi-word lexical items such as separable verbs have been counted as one word; the target word (a common word by default) has been included in the common word count for every version. Multiple word matches were not allowed. More sophisticated measures of the lexical overlap method were not considered, as they would have required the use of sources that are either computationally costly (e.g. corpus-based information for the use of weighting factors based on information gain), or available only to a very limited extent (information on lexical relations such as hyponyms in the lexicographic resources listed above).

Two frequently used similarity measures for common element-based overlap methods, the Dice and the Jaccard coefficient, have been used to derive the sentence similarity scores. The Dice coefficient is defined as $\frac{2|X \cap Y|}{|X|+|Y|}$, i.e. it normalizes for length.[23] The Jaccard coefficient is defined as $\frac{|X \cap Y|}{|X \cup Y|}$ and has the effect of penalizing a small number of shared entries more than the Dice coefficient does (Manning and Schütze, 1999, p. 299). Both measures range from 0 (no overlap) to 1 (perfect overlap).

[22] an online dictionary available at `http://wortschatz.informatik.uni-leipzig.de/index_js.html`

[23] It is equivalent to the F-measure $\frac{2PR}{P+R}$, where Precision $P = \frac{Common-elements}{Elements-in-S_1}$ and Recall $R = \frac{Common-elements}{Elements-in-S_2}$.

5.4.1 Results of Analysis

The results of the correlation analysis between the 2x4 versions of lexical overlap and the human ratings are provided in table 5.3.

Table 5.3: Correlations of Sentence Similarity Lexical Overlap Measures with Average Human Ratings

$^{**}/^{*}$ Correlation is significant at the 0.01/0.05 level (2-tailed)

Lexical Overlap Version	**Measure used**	***r***
Basic Lexical Overlap	Dice	0.45**
Basic Lexical Overlap	Jaccard	0.45**
Content words-only Overlap	Dice	0.29
Content words-only Overlap	Jaccard	0.27
Basic + Synonym Overlap	Dice	0.22
Basic + Synonym Overlap	Jaccard	0.15
Content + Synonym Overlap	Dice	0.44**
Content + Synonym Overlap	Jaccard	0.44**

5.4.2 Discussion

Somewhat surprisingly, the most basic version of lexical overlap which includes function words such as pronouns and prepositions in the overlap count performs best with a moderate correlation to the human rating averages of $r = 0.45$ (both for the Dice and Jaccard coefficients). The basic version performs marginally better than the most complex version (combining synonym enhancement with the exclusion of stopwords), which also achieved an — almost as high — moderate correlation of $r = 0.44$. Surprisingly, both the exclusion of stopwords and the consideration of synonyms on their own lowered the performance of lexical overlap to below-moderate, non-significant correlation levels of $r < 0.30$. The correlations found for basic lexical overlap are on a similar level as the ones reported by Wiemer-Hastings (1999).[24]

While the data set tested is too small to speculate on possible reasons for the poor performance of content words-only and synonym-enhanced lexical overlap measures,

[24]The results are not directly comparable as Wiemer-Hastings correlated the measure at a range of different threshold levels, and collected the human ratings by asking the raters to say how much of a student answer matched the (longer) expected answer.

it should be noted that the synonym-enhanced version used a very conservative criterion for synonym detection (see above); not only did it not cover other lexical relations possibly employed by teachers (antonyms, hypernyms), but it also failed to consider many instances of 'phrasal synonymy', i.e. paraphrasing, as in *Briten* (britons) vs *die Bewohner des Vereinigten Königreichs* (U.K. residents). Increasing the penalty for low-overlap pairs (via the Jaccard coefficient) had no beneficial effect in terms of the correlations achieved.

5.5 Analysis of Sentence Similarity Teacher Data with Latent Semantic Analysis

The training corpus used for the LSA analysis was the *Frankfurter Rundschau* (FR) corpus, which collects all editions of the daily German newspaper between ca. June 1992 and March 1993, and has a total size of ca. 230 MB. Since both corpus size and dimensionality are assumed to have an effect on LSA's performance (see section 5.2.4.1), the versions of LSA analyzed were varied along both of these dimensions. The tested training corpus sizes range from 2.3 MB (ca. 17,000 terms) to 150 MB (ca. 250,000 terms) — also depending on the corpus type, see below. The tested sizes are thus substantially higher than the training text sizes used in similar studies of LSA *vs* human sentence similarity ratings (the lower limit of ca. 2 MB corresponds roughly to the size of the training corpus used by Wiemer-Hastings et al. (1999); Zipitria et al. (2006)). The full size of the FR corpus could not be used due to the computational complexity of the SVD compression exceeding the capacity of available computational resources for that size.

The dimensionality was varied in steps of 50 dimensions (D) ranging from 200 D to ca. 500 D (depending on the corpus size).

Five different versions of the FR corpus were analyzed to investigate the benefits of (a) sanitizing the corpus, i.e. removing non-sentential 'junk data' such as event listings, sport results, addresses etc., and (b) varying the document size between paragraphs and entire newspaper articles. The latter is the default for the FR corpus; however, the use of smaller paragraph-sized documents seemed advisable as it has been suggested that paragraph-sized documents might be beneficial for the performance of LSA due to the arguably increased "topical" coherence of the paragraphs as opposed to longer text segments (Wiemer-Hastings et al., 1999).

As is customary in LSA applications, a stopword list (in this case consisting of the 200 most frequent word forms in German) was used to filter out frequent function words arguably contributing little to the sentence meaning, thereby reducing the computational load thanks to the smaller matrix size. The bulk of the training corpus consisted of varying-sized chunks of the FR corpus ($> 99\%$ of the entire training corpus), plus the selected 40 sentence pairs (see section 5.3) and all 17 reading texts given to the teachers in the exploratory study.[25]

The following five versions of the FR corpus have been analyzed:[26]

- **Corpus-A**: LSA documents are entire articles; all 'junk data' left in (no sanitizing);
- **Corpus-B**: LSA documents are entire articles; 'junk data' almost completely removed ($\sim 100\%$);
- **Corpus-C**: LSA documents are entire articles; 'junk data' mostly removed ($\sim 90\%$);
- **Corpus-D**: LSA documents are article paragraphs; 'junk data' left in (no sanitizing);
- **Corpus-E**: LSA documents are article paragraphs; 'junk data' almost completely removed ($\sim 100\%$).

5.5.1 Analysis Results

The results of the correlation analysis of the different combinations of corpus type, size and dimensionality are given in table 5.4. The correlation analysis was based on a correlation of 'raw' cosines with the human ratings, i.e. no correction was applied for occasionally occurring negative cosines.[27] The cosine measure was chosen, as it is the

[25] It was subsequently discovered that the human-rated sentence pairs given to LSA contained some typos, most of which had no effect by default as they concerned double appearances of stopwords; one typo however affected a content adjective. The correlation analysis was re-run for the best-performing LSA/corpus version (see below), with no difference in the overall correlation found up to the third decimal place. It can therefore be assumed that the effect of the typos on the overall result was negligibly small.

[26] The reason that both Corpus-B and Corpus-C were used despite the relatively small difference in sanitizing was that Corpus-C was used first as the 'pilot' training corpus; Corpus-B was developed later to investigate if full sanitizing was beneficial.

[27] In theory, negative cosines should not occur, because of all of the original data consist of non-negative numbers (weighted word counts). However, they do occur in practice because the reduced k-dimensional representation is just an approximation of the original data (Wiemer-Hastings, personal

Table 5.4: Correlations of LSA cosines with average human ratings of Sentence Similarity

Corpus-A (docs are articles; no sanitizing)	**200D**	**250D**	**300D**	**350D**	**400D**	**450D**	**500D**	**550D**	**600D**	**1000D**	**1400D**	**1600D**
73 MB (156,477 terms)	0.63	0.67	0.67	0.67	0.72	0.72	0.71					
135 MB (216,444 terms)	0.56	0.61	0.63	0.68	0.67	0.67	0.67	0.67	0.68			
Corpus-B (docs are articles; full sanitizing)												
42 MB (120,495 terms)				0.64	0.65	0.65	0.65					
67 MB (157,555 terms)				0.73	0.75	0.74	0.73					
Corpus-C (docs are articles; 90% sanitizing)												
2.3 MB (17,163 terms)	0.60		0.62	0.62	0.60	0.60	0.60	0.62	0.60			
6 MB (34,687 terms)			0.53	0.54	0.57	0.58	0.58	0.60	0.60	0.63	0.65	0.59
19 MB (73,848 terms)			0.54	0.58	0.60	0.58						
31 MB (101,007 terms)			0.63	0.63	0.64	0.65	0.64	0.62				
44 MB (123,572 terms)			0.69	0.71	0.70	0.71	0.69	0.68				
56 MB (142,904 terms)			0.72	0.75	0.73	0.71						
68 MB (159,859 terms)	0.62	0.66	0.73	0.75	0.75	0.74	0.73					
87 MB (183,601 terms)				0.74	0.75	0.75	0.73					
126 MB (225,043 terms)	0.66	0.69	0.71	0.74	0.74	0.74	0.75	0.75	0.68			
150 MB (249,932 terms)	0.66	0.67	0.70	0.70	0.71	0.72	0.73					
Corpus-D (docs are paragraphs; no sanitizing)												
74 MB (165,625 terms)	0.65	0.63	0.63	0.58	0.58	0.58	0.59					
124 MB (228,956 terms)	0.57	0.58	0.58	0.52								
137 MB (243,020 terms)	0.57	0.58	0.57									
Corpus-E (docs are paragraphs; full sanitizing)												
67 MB (158,927 terms)	0.64	0.66	0.65	0.62	0.58	0.60	0.60	0.59				
124 MB (234,107 terms)	0.62	0.63	0.58	0.53	0.53							

most widely used and easily obtainable similarity measure in LSA and vector-model applications.

Small corpus sizes (down to 2.3 MB) were only analyzed for the 'pilot' Corpus-C, as Corpus-B yielded very similar maximum correlations as Corpus-C, and the other corpora did not achieve as high peak correlations as Corpus-B/C. Dimensions above 500 D were only analyzed if it was computationally feasible, and it had not already become apparent from the lower-dimensional results for the given corpus type/size combination that the peak correlation occured below 500 D. For the same reason, for some corpus type/size combinations, lower dimensions (below ca. 300 D) were not analyzed. All correlations listed in the table are significant at $p < 0.01$ (two-tailed).

5.5.2 Discussion

The correlation analysis revealed that the best-performing LSA/corpus combination was for Corpus-B (documents are articles; full sanitization), for a corpus size of ca. 67 MB with 350 dimensions, $r = 0.75$; several other versions for the very similar corpora B and C performed on nearly the same level.

These correlation results are significant as they not only show that LSA (for German and with a corpus with similar characteristics as the FR corpus) achieves a high correlation with human ratings that clearly outperforms at least basic versions of the lexical overlap approach, but also since these results lend further support to the hypothesis suggested by the findings of Zipitria et al. (2006), namely that LSA can perform the task of rating the similarity of sentence pairs considerably better when applied to languages other than English that provide more syntactic clues to LSA due to their higher position on the synthetic/agglutinative language scale. Zipitria et al. reported peak correlations of $r = 0.71$ for lemmatized Basque and $r = 0.61$ for non-lemmatized Spanish. To the author's knowledge, these results are also the highest correlations reported so far for lemmatized and standard, non-lemmatized version of LSA, respectively. The correlation levels achieved in this study at least match the results for lemmatized Basque, and clearly exceed previously reported correlation levels for standard, non-lemmatized LSA. As a slight caveat, it should be noted that (a) most studies of a similar nature that have compared either basic or syntax-enhanced versions of LSA to human ratings (Wiemer-Hastings, 2000; Wiemer-Hastings and Zipitria, 2001; Kanejiya et al., 2003; Wiemer-Hastings, 2004; Zipitria et al., 2006) are not always directly comparable be-

communication). Since the occurring negative cosines are always very small (down to ca. cos=-0.10), their effect can be assumed to be similar to zero cosines.

cause of the different way the human ratings were obtained and the correlation analysis was conducted (see section 5.4), and (b) that the human ratings of the study at hand were based on a rather limited set of 40 sentence pairs.

On the basis of these results, no definite conclusion can be drawn on the extent of the relative benefits of the two main differences of LSA usage to previous, similar studies, namely the considerably increased corpus size, and the use of a synthetic language like German that, thanks to its comparative morphological richness, provides LSA with more syntactic-relational "clues" than English does. However, the results do indicate that both factors have contributed to the considerably improved performance: first, for the smallest corpus size of 2.3 MB (similar to the size used in similar studies of Wiemer-Hastings (1999); Zipitria et al. (2006)), the correlations achieved were in the ballpark of $r = 0.60$, which is on a similar level as the correlations for non-lemmatized LSA for Spanish by Zipitria et al., and notably higher than the peak correlation of ca. $r = 0.48$ found by Wiemer-Hastings for English. Second, the results, in particular for Corpus-C, show that an increased corpus size (up to ca. 70 MB) leads to an increase (albeit a non-linear one) in performance.[28] Slightly surprising, however, is the finding that for further increases in corpus size (up to ca. 150 MB), the performance gain not only diminishes, but LSA's overall performance actually tends to worsen.

With respect to the different corpus types investigated, it appears that sanitizing the corpus has a moderately beneficial effect at least for documents-as-articles corpora, with a peak correlation of $r = 0.75$ compared to the non-sanitized peak of $r = 0.72$. This finding is surprising as sanitizing the corpus has been assumed to yield no noteworthy payoff (see section 5.2.4.1). With respect to the granularity of documents, even though paragraph-level is the generally recommended size for LSA documents, the finding that the performance level of documents-as-articles versions is considerably higher in this study is perhaps not so surprising as it may appear at first glance. This is because many of the paragraphs in newspaper texts such as the FR corpus are very short, and articles tend to be more focussed in newspaper texts than in they would be in other text sources (e.g. book chapters).

It remains to be seen in future research whether lemmatization or the addition of structural information to LSA can yield a further improvement in performance for German.

[28] This result is in keeping with previous findings on the effects of corpus size (Wiemer-Hastings et al., 1999; Landauer, 2002).

5.5.3 Summary

This section has established that LSA achieves high correlations with human sentence similarity ratings and significantly outperforms basic lexical overlap as an alternative measure. It has therefore been chosen as the measure of sentence similarity for the remainder of this thesis, using the version of LSA that achieved the best correlation with the human ratings (Corpus-B, 67 MB, 350 D). In the next section, LSA will be used to establish whether sentence similarity is a significant criterion in the teacher data.

5.6 Significance Analysis of Sentence Similarity in the Teacher Data

Having shown the validity of LSA as a measure of sentence similarity, it remained to be established whether sentence similarity is a significant criterion in the selection of teacher examples. This issue has been approached in the following way: first, a precision-recall analysis was conducted on the human ratings and LSA cosine data to determine which partition of the cosine range yielded the highest F-score, i.e. best predicted the human ratings with a minimum of false positives and negatives, and the corresponding cosine threshold separating the highest partition (containing high similarity items) was determined. Second, based on the above partition and threshold, the ratio of sentence pairs in the teacher data classified as 'highly similar' was computed. Third, sentence pairs were selected at random from corpora, and their ratio of highly similar pairs was compared with the ratio in the teacher data, in order to test the hypothesis that the percentage of highly-similar sentence pairs in the teacher data is significantly higher compared to random pairs that share the same target word. The following subsections describe these steps in turn.

5.6.1 Determining High-Similarity Thresholds via Precision-Recall Analysis

In order to find out which partition of the cosine range best predicted the human ratings, a precision-recall analysis was carried out on LSA cosines and corresponding human

ratings averages for 31 out of the total 40[29] test sentence pairs, with
$Precision = \frac{correctly-predicted-elements-in-partition}{predicted-elements-in-partition}$ and
$Recall = \frac{correctly-predicted-elements-in-partition}{elements-in-partition}$.

The following algorithm was used to determine the optimal cosine threshold for a possible number of *n* partitions, with *n* ranging from 2 to 4. Taking *n*=2 as an example, first the human rating scale was divided into 2 equal-sized partitions, with *elements-in-partition* being the number of sentence pairs belonging to the respective partition. Possible thresholds in the cosine space were then increased or decreased in steps of cos=0.05, with *predicted-elements-in-partition* corresponding to the number of sentence pairs above or below the threshold in question. The 'winning' threshold was taken to be the threshold with the highest F-score (for higher numbers of *n*, the final F-score was computed as the average F-score after appropriate mergings of partitions).

The highest F-score (F=0.85) in this precision-recall analysis was found for just two partitions (high and low similarity) for a threshold of cos=0.35.[30]

The threshold of cos=0.35 was then used to determine the percentage of high-similarity sentence pairs in the teacher data, with 22 sentence pairs (9.1% of the total items) found to be above that threshold. Since this analysis was based on the complete set of teacher sentence pairs (i.e. *not* resolved for anaphora), but the LSA analyses up to this point had been based on the test set of 40 sentence pairs that *had* been resolved for anaphora, a second, anaphora-resolved version of the complete teacher data set was used, resulting in a re-classification of 1 sentence pair from high to low similarity (overall ratio of highly similar items for this version: 8.6%). This slightly more conservative figure was then used as the yardstick against which the percentage of highly similar *random* sentence pairs was compared.

5.6.2 Selection of Random Sentence Pairs from Corpora

A test set of 250 random sentence pairs was derived in the following way: first, 25 target words were selected from the total set of target words. To achieve this, all 243 target words were ordered in descending order of their associated sentence pair similarity (as measured by the LSA cosine). They were then split into 25 groups containing

[29] Since the to-be-determined cosine threshold was going to be used for the analysis of the complete set of teacher sentence pairs, the precision-recall analysis was restricted to the 31 human-rated sentence pairs which had been retained in the final teacher data set (see section 5.3.3).

[30] Since none of the 31 test sentence pairs had associated LSA cosines between 0.35 and 0.40, the data did not allow to distinguish between these two thresholds, and the cos=0.35 threshold was chosen arbitrarily.

mostly 10 words each, with one word being selected from each group such that the overall part-of-speech ratio[31] and word frequency distribution[32] of the entire set of target words were roughly preserved.

Second, for each of the 25 word selections, 10 sentence pairs were selected at random from 3 different corpus sources in the following way:

- 3 sentences from *Wortschatz Leipzig*[33] (an online dictionary with (depending on word frequency) up to ca. 250 authentic example sentences from various newspaper and literature sources for each entry);

- 3 sentences from the *DWDS* corpus[34] (selections only taken from source period 1950-2005);

- 4 sentences from the *IDS Mannheim* corpus[35].

A random sentence was discarded and replaced with a new random sentence for any of the following reasons:

- The target word was used in different sense than in the original sentence. This decision was made based on instrospection of the experimenter, using word sense entries in *DWDS* as a guideline. Differences in literal *vs* figurative usage were not considered as a filter criterion;

- It was obvious the sentence was used in a literary or historical context;

- The sentence was incomplete or contained at least one idiosyncratic abbreviation (e.g. B. for *Buddhismus* (buddhism));

- The target word was a name or title (e.g. of a literary work).

The result of the selection was a set of 250 random sentences, the cosines of which with their corresponding original sentences were then computed. The results of this analysis are presented in the next subsection.

[31] The 25 selected words consisted of 12 nouns, 6 verbs, and 7 adjectives/adverbs.

[32] as measured by the IDS count for lemmas; the words were ordered by this frequency count and evenly split into 5 groups. 5 words were then selected from each of these groups.

[33] available online at `http://wortschatz.informatik.uni-leipzig.de/index_js.html`

[34] available online at `http://www.dwds.de`

[35] available online at `http://www.ids-mannheim.de/cosmas2/`

5.6.3 Determining Sentence Similarity for Random Sentence Pairs

The cosines obtained for the random sentence pairs described above had a mean of cos=0.08, St.Dev.=0.08 (n=250); this compares to a mean of cos=0.13, St.Dev.=0.15 (n=243) for the teacher data cosines.[36]

Both parametric and non-parametric tests (t-test and significance analysis based on the classification data) were conducted to investigate the following related but distinct questions:

- Do teachers, on average, select examples that are significantly *more* similar to their corresponding original sentences than what would be expected in random sentence pairs?

- Do teachers use significantly more examples that can be considered *highly* similar to their corresponding original sentences than what would be expected in random sentence pairs?

To address the first question, an independent-sample[37] t-test was conducted on the teacher and random cosine data. The t-test revealed that, on average, the similarity of the teacher-selected sentence pairs, as measured by the LSA cosines, is significantly higher than that of randomly selected sentence pairs ($t(378,851) = -4.906$, $p < .01$).

To address the second question, a classification analysis was carried out using the high similarity threshold of cos=0.35 derived in the previous section. The classification for both random and teacher sentence pairs is reported in table 5.5. As can be seen from the table, the ratio of highly similar sentence pairs is considerably lower for random sentence pairs (1.2%) compared to the teacher data (8.6%).

Table 5.5: Similarity Classifications for Teacher and Random Sentence Pairs

Classification	**Teacher Data** (n=243)	**Random Data** (n=250)
High Similarity	21 (8.6%)	3 (1.2%)
Low Similarity	222 (91.4%)	247 (98.2%)

[36]The mean found for the random sentence pairs is higher than the one reported by Landauer (2002) for random passage-to-passage cosines (0.04).

[37]The independent-sample t-test was chosen because teacher cosines do not appear to impact random cosines and vice versa; however, this choice is slightly problematic as the target words used for the random pair analysis are a subset of the teacher data target words.

To find out whether this difference of relative occurrences of highly similar sentence pairs is significant, the Pearson chi-square (χ^2) test has been conducted.

The χ^2 test revealed a significant difference ($\chi^2(1)$=14.74, $p < .01$) in similarity classifications for teacher and random sentence pairs.

In sum, these results indicate that, on average, teachers use examples that are significantly more similar to their corresponding original sentences than what would be expected from a random selection, and that significantly more examples can be considered highly similar to their corresponding original sentences than what would be expected from random sentence pairs. This means that even though the ratio of highly similar sentence pairs in the teacher data is relatively small (ca. 9%), it still is significantly higher than what could be expected by chance. In sum, sentence similarity has been found to be a significant factor in the selection of the teacher examples, and will therefore be included in the regression model of the teacher data (see chapter 7) as one of the predictor variables.

5.7 Summary

This chapter addressed the question of whether sentence similarity is a significant factor in the choice of teacher examples; more specifically, it presented an analysis on whether the degree of sentence similarity of the sentence pairs in the teacher data is significantly higher than that found in randomly collected pairs. This question has been answered in the affirmative, motivating the inclusion of sentence similarity as a predictor variable to be included in model of teacher criteria for the selection of examples (see chapter 7).

As a pre-requisite first step to addressing the above problem, several potential measures of sentence similarity were surveyed. Of the two methods selected for analysis — lexical overlap and LSA — LSA was found to yield the best correlation to human sentence similarity judgments, and was therefore used as the measure of choice in the analysis of the teacher and random sentence pair data.

Chapter 6

Specific Lexical Choices in the Teacher Examples

6.1 Introduction

In chapters 4 and 5, we looked at the teachers' examples from a 'global', sentence-level point of view, both from a syntactic (complexity) and semantic (similarity) perspective. In the current chapter, the focus of attention will be narrowed to a 'local' point of view, namely an investigation into what kinds of *lexical* choices were significant criteria for the teachers in their selection of example sentences.

The analysis of the teachers' explanations in chapter 3 suggested that teachers made specific lexical choices with regards to specific words in their examples that are *semantically related* to the target word, in order to provide a lexical clue to the meaning of the target word, as well as to the *morphological form* of the target word itself.

Before these categories are described in further detail, the use of the terms *semantic relatedness* and *(semantic) word similarity*[1] should be clarified first, as they are very broad notions the usage of which is not always clearly defined in the literature. The term *word similarity* is used in this chapter in the more general sense of semantic relatedness rather than strict semantic similarity which, as Budanitsky and Hirst (2001) have pointed out, would imply a restriction to synonymy-type relations. Semantic relatedness, however, includes not only other paradigmatic lexical relations such as hyponymy and antonymy, but also 'topical' relations of a syntagmatic kind which are

[1] Throughout this chapter, *word similarity* will be used in the sense of *semantic* word similarity, i.e. it refers not to the morphological similarity of word forms, but rather to that of the respective word senses or concepts they denote.

often indicated by significant co-occurrences (e.g. *doctor* and *nurse*). As teachers have indicated both types of lexical choices in their explanations (see chapter 3), it is this broader category of semantic (also known as sense or lexical) relations that are of interest in the context of this study.

Paradigmatic relations exist between lexical units that share at least one semantic core component, belong to the same part-of-speech, and fill in the same syntactic position in a syntactic construction. The most common paradigmatic relations are synonymy, antonymy, and hyponymy/hyperonymy. Synonymy in particular is a matter of degree rather than an absolute concept; a strict interpretation requires that for two words to be considered synonymous, they would have to be mutually substitutable in *every* context. However, this requirement is too stringent to be met in practice by the vast majority of synonymy candidates, so often a more lenient criterion in the form of partial or quasi-synonymy is adopted by dictionaries and lexicographers.[2]

Similar comments regarding the gradability of the concepts apply to the other main paradigmatic relations — antonymy and hyponymy/hyperonymy — which are preferentially found in specific parts-of-speech: antonymy, which can be viewed as an association between two lexical units which have the opposite core meanings in some contexts, is most commonly encountered among adjectives. In his analysis of vocabulary elaboration in teachers' L2 classroom language, Chaudron (1982) observed that "opposites are probably very noticeable for the L2 learner, owing to the predominant use of negation in such elaboration [...]". By way of contrast, hyponyms/hyperonyms (subordinates/superordinates), are typically met in the noun and verb categories.[3]

For the purposes of this chapter, the range of paradigmatic lexical relations taken into consideration has been limited to the most commonly found types mentioned above. The following operational criterion of what is considered a paradigmatic relation to the target word has been adopted: a word is considered to be in a lexical relation[4] to the target word if it is listed as such in one of the relevant lexicographic resources mentioned in section 5.4.[5] For polysemous words and homonyms, this leaves

[2]Partial synonymy only requires substitutability in *some* context, or in Cruse's (1986) terms, that the respective word senses are identical in their central semantic traits, but may differ in minor or peripheral traits, in which case it would be up to the lexicographer to set the threshold below which differences are considered permissible.

[3]For verbs, sometimes the distinction between *entailment* (e.g. *to snore* entails *to sleep*) and *troponomy* (*to march* is a troponym of *to walk*) is made (Fellbaum, 1998).

[4]For the remainder of this chapter, the term 'lexical relation' will be taken to mean 'paradigmatic lexical relation' only.

[5]GermaNet (the German version of WordNet), *Wortschatz Universität Leipzig* (`http://wortschatz.informatik.uni-leipzig.de/index_js.html`), and the monolingual dictionaries

the problem of word sense disambiguation (WSD) in order to identify the correct sense in which the word is used in the given context. Section 6.2 elaborates how the WSD problem has been approached for the current study.

Having obtained the 'gold standard' for identifying paradigmatic lexical relations in the teacher data in this way, the goal in section 6.3 is twofold: first, to investigate whether this manually derived gold standard of (paradigmatic) word similarity can be replaced by word similarity measures proposed in the literature; the validity of the selected measures will be tested both in a TOEFL[6]-type multiple-choice lexical-relation identification test and against human similarity ratings of word pairs. Second, using the thus selected measure of word similarity, the teacher examples will be analyzed in terms of their use of words that are paradigmatic relations to the target word.

In this chapter, *syntagmatic* relations (to the target word) of interest are taken to be collocational relationships in the sense of *significant co-occurrences*[7] of the target word. For teachers, these can be presumed attractive as possible lexical choices in their examples for two reasons: first, collocational relationships appear to have "powerful and long-lasting" links in the mental lexicon (Aitchison, 1994, p. 90); cutting across part-of-speech boundaries, they often provide associative clues to its meaning (e.g. for the target words *to bark* and *dark*, a teacher might use an example containing the co-occurrences words *dog* and *night*, respectively). Second, the use of significant co-occurrences provides important clues as to the *usage* of the target word, which is widely considered to be a part of the dual function of example sentences (see the discussion in chapter 2). Even though the teachers had been told the primary consideration for their selection of examples should be the other part of that dual function, namely the *illustration of the meaning* of the target word, it can thus be speculated that this usage-related function played a role in their selection of examples at least to some extent. The analysis of teacher explanations in chapter 3 suggested that syntagmatic lexical associations play an even more important role within the general category of specific lexical choices than paradigmatic relations do.

The goal of section 6.4 in this chapter is to analyze the teacher data (based on corpus co-occurrence statistics) in terms of whether the use of words that are significant

---

WAHRIG, *DUDEN Bedeutungswörterbuch*, *PONS Großwörterbuch*, and *Langenscheidt Großwörterbuch Deutsch als Fremdsprache*.

[6]Test of English as a Foreign Language

[7]Throughout this chapter, Manning and Schütze's (1999) distinction between the terms *co-occurrence* and *collocation* is adopted: the latter denotes "grammatically bound elements that occur in a particular order", whereas *co-occurrence* refers to "the more general phenomenon of words that are likely to be used in the same context" (p. 185).

co-occurrences of the target word plays a more significant role in the teacher examples (ES) compared to the original sentences (OS). It should be noted at this point that significant co-occurrences and lexical relations of the target word are not mutually exclusive; in the analysis to follow, words in the examples that meet both criteria are counted in both categories.

The second main category of lexical choices pertains to the *morphological form* of the target word, which — depending on the respective part-of-speech category — may have been chosen by the teacher (e.g. for simplification purposes), in order to highlight the most frequently used form, or to indicate the gender of a noun.

The chapter is organized as follows: section 6.2 explains how the problem of WSD is handled in the remainder of this thesis. Section 6.3 first provides an overview and discussion of measures of word similarity, especially with a view to their suitability as measures for the current study, then proceeds to justify the use of a dictionary-based 'manual' gold standard as the measure of choice, and concludes by presenting the analysis of the teacher data in terms of common lexical relations. Section 6.4 deals with the analysis of significant co-occurrences in the teachers' examples, and section 6.5 presents the analysis of the specific morphological forms of the target words used in the examples. Finally, section 6.6 concludes with a summary of the chapter.

6.2 Word Sense Disambiguation

The problem of WSD is outside the scope of this thesis and has therefore been dealt with on a manual basis where necessary. A dictionary-based approach would consist of (manually) identifying the appropriate sense[8] in which a given target word is used in the OS, and then selecting appropriate lexical relations or examples containing the target word in that sense. This approach had been considered but was rejected for the following reasons: (a) available current dictionaries tend to vary in the number of different senses they assign to polysemous words, so that any such standard would ultimately be arbitrary; (b) wide-coverage dictionaries in particular, as well as conceptual networks such as *GermaNet*, often employ sense distinctions that tend to be too fine-grained for the purposes of this study: while a cursory inspection of the teacher data seems to bear out the intuitive assumption that teachers would avoid using the target word in a clearly distinct and unrelated sense (as is the case with homonyms such as *bank*), teachers cannot be assumed to restrict their examples on the basis of

[8] usually indicated by a separately numbered subentry in a word's main dictionary entry

sense distinctions that are too elaborate for this purpose. This is especially the case considering that in some cases, sense distinctions are too fine-grained to even permit an unequivocal 'manual' assignment of a given data item to a given word sense.

A related problem concerns the distinction between *literal* and *figurative* usages of a word. Intuition suggests that teachers would employ the word only in a literal usage in their examples unless the OS had used the word in a figurative sense, in which case either usage could be imagined in the example; after all, using an unknown target word figuratively in the ES would introduce an additional level of difficulty in case the OS contained a literal usage of that word. An inspection of the teacher data corrobrates this intuition-based hypothesis: while there are no transitions from literal to figurative usage from OS to ES, the ratio is roughly 60:40 in favor of the 'figurative-to-literal' transitions *vs* 'figurative-to-figurative' ones. Therefore, the manual WSD selection adopts the above hypothesis on the issue of figurative *vs* literal usage.[9]

The above being the case, the adopted manual approach to WSD was a liberal one and consisted of allowing all closely related senses for a given word that seemed plausible candidates for a teacher selection, based on introspection of the author and aided by dictionary entries as a guideline. The respective dictionary entries were consulted to either (a) confirm the author's intuitions regarding the closeness or remoteness of two senses or (b) in dubious cases, provide clarification of the 'closeness' of the two senses. The information of interest to be gleaned from the dictionaries was whether the appropriate word senses were listed as separate entries or subentries, plus any potential clarifications sense glosses could provide if they were available.

While it has to be conceded that the resulting decisions are, to some extent, arbitrary, especially as there is no clear dividing line between 'closely related' and 'remotely related' senses of a word, it is also noteworthy that a teacher would have to make the same informed but somewhat arbitrary decisions in his example selections, and usually *without* the aid of lexicographic resources. Furthermore, for the given application of advanced-level vocabulary learning, the WSD problem is more one of principle than of practice, since target words on this level tend to be infrequent words, and as a general rule of thumb, polysemy increases with word frequency.

For the lexical relation identification task relevant to this chapter, this means that a word (sense) A is considered a lexical relation to word (sense) B if the lexical relation is included in the union of lexical relations listed for all appropriate dictionary word

[9] While the literal *vs* figurative usage distinction is of no relevance for the analyses in this chapter, it will be an issue for the example selection in chapters 7 and 8.

sense entries. For the example selection task in chapters 7 and 8, it means that every sentence containing the word in one of the appropriate dictionary word senses is a suitable example sentence candidate.

6.3 Measuring Word Similarity

This section examines the issue of whether, for the purposes of this study, the manual gold standard of measuring word similarity by a count of common lexical relations (on the basis of suitable dictionary and conceptual network listings) can be replaced by a suitable automated measure of word similarity.

To this end, first an overview will be given of measures of word similarity proposed in the literature, with a view to their suitability for the analysis task at hand. The selected measures will then be compared in two commonly used word similarity tasks: (a) a TOEFL-style multiple-choice lexical relation test, and (b) a correlation analysis with human word similarity judgments.

6.3.1 Measures of Word Similarity

Since the task of measuring word similarity can be seen as a special instance of the general task of measuring text similarity, which has been discussed in chapter 5 in the context of sentence similarity measures, the overview on measures of word similarity in this section will be confined to aspects particular to word similarity not covered in the previous discussion.

6.3.1.1 Dictionary or Thesaurus-based Overlap Measures

Arguably the most basic group of word similarity measures is based on an overlap count of common (word) elements in the dictionary (or other WordNet-style conceptual network) glosses of word senses based on the Lesk algorithm (Lesk, 1986; Banerjee and Pedersen, 2003). This kind of approach has not been considered for this analysis because (a) it would have required the creation of 'artificial' glosses due to the insufficient availability of glosses in available lexicographic resources for the set of target words; (b) for German, the method has been found to perform on a lower level than alternative taxonomy-based measures in terms of correlations to human judgments of semantic similarity (Gurevych and Niederlich, 2005).

6.3.1.2 Semantic Network-based Measures

A common trait of measures in this group is that they are based on a lexical resource in the form of a WordNet-type conceptual network or directed graph, and compute word similarity on the basis of properties of the network structure, or type and properties of the network paths (e.g. path length and/or direction, relative depth and density). A thorough overview of these measures can be found in Budanitsky and Hirst (2006), so suffice it to say at this point that two main subgroups can be distinguished among them: (a) measures that are purely taxonomy-based in the way just described (e.g. Hirst and St-Onge (1998); Jarmasz and Szpakowicz (2003)), and (b) hybrid measures that augment the taxonomy-based measure with some form of corpus statistics-derived information content (e.g. Resnik (1995); Jiang and Conrath (1997); Lin (1998); Li et al. (2003)). The idea behind these hybrid measures is that the similarity of two concepts is related to their shared information content (as indicated by e.g. a highly specific common subsuming concept), which in turn is directly related to the frequency of occurrence of the corresponding terms in a corpus. Taxonomy-based measures (both of the pure and hybrid variety) have not been considered as word similarity measures for this study for the same reasons they have been excluded from consideration as sentence similarity measures (see chapter 5).[10] For purposes of word similarity ratings, in addition to its relatively low coverage, WordNet's German equivalent, GermaNet, has the disadvantage of including artificial, non-lexicalized concepts due to its design principles being based on linguistic evidence rather than psycholinguistic motivations (Gurevych and Niederlich, 2005).

6.3.1.3 Distributional Similarity Measures based on Syntactic Context

Distributional similarity measures based on syntactic context derive their similarity scores of two words on the basis of their distribution in a text corpus. Their scores are based on some information-theoretic measure applied on the basis of a grammatical analysis of the parsed texts. For example, the underlying assumption of Lin's (1998) or Grefenstette's (1994) approach is that synonyms tend to be found in similar grammatical frames (this has been dubbed the 'parallelism' assumption by Higgins (2005)); the basic data for e.g. Lin's similarity score are 'dependency triples' consisting of a

[10] An additional reason for not using hybrid measures (which have generally been found to outperform 'pure' taxonomy approaches) for German, is that the complex morphological structure (in comparison to English) would have necessitated the employment of a morphological analysis component more complex than stemming in order to achieve accurate mappings from word frequency counts to word senses (Gurevych and Niederlich, 2005).

word pair and the grammatical function relating them[11].

Measures of this type have not been considered for the analysis, as they would require extensive parsing and be computationally too expensive for the application at hand.

6.3.1.4 Probabilistic Measures

These measures are based on distributional information in the form of probability distributions and are usually cast as measures of dissimilarity between these distributions (Dagan et al., 1997). These measures have not been considered for the analysis either due to their relative computational expensiveness.[12]

6.3.1.5 Statistical Vector-Space Measures

Vector-based measures in general, and LSA in particular, have been covered in detail in chapter 5 in relation to sentence similarity. The following discussion will therefore be confined to listing the reasons why LSA has been selected as a word similarity measure to be tested further in the remainder of this section:

- LSA is the most widely cited among vector space-type measures, especially in relation to word similarity judgments, and has been shown to perform well as a word similarity measure in TOEFL-style synonymy judgment tasks (Landauer and Dumais (1997) have reported an average LSA score of 64% correct answers for the synonym-section of the TOEFL test which, as they note, would have been "adequate for admission to many universities");

- In addition to an attested good performance in synonymy tasks, LSA has demonstrated its ability to detect the semantic relatedness of antonyms and morphologically related words (Landauer, 2002);

- LSA has been shown to perform well as a sentence similarity measure for the current study (see chapter 5), and has been selected as an analysis tool for the teacher data for this purpose; as has been noted in the previous chapter, LSA's performance in word similarity tasks has generally been found superior to its performance in the task of sentence similarity ratings;

[11] For example, the dependency triples for the sentence "I have a cat" would be (have subj I), (have obj cat), (cat det a).

[12] Probabilistic dissimilarity measures also require an additional transformation to derive a measure that can be directly used for nearest neighbor generalization (Manning and Schütze, 1999).

- Being based on the 'topicality' assumption (synonyms tend to have the same neighbors since they tend to occur in passages which are on same topic), LSA can capture topical similarity, as well as paradigmatic relations other than synonymy. This ability to capture all sorts of semantic relatedness may not necessarily be an advantage for this analyis, as LSA does not distinguish between different types of relatedness. However, the same observation also applies, in varying degrees, to other similarity measures as well.

The following slight caveats regarding the use of a vector-space measure such as LSA for the task of German word similarity judgments should be noted at this point: first, despite the general above-mentioned success of LSA in word similarity rating tasks, the observed performances in similarity rating tasks are less than optimal, most likely due to the sparse data problem afflicting all corpus-based measures and the non-recognition of lexical ambiguity — multiple senses of a word are "lumped together" (McDonald, 1997). Second, most of the the corresponding research has been done for English; for a morphologically richer language such as German, it is not *a priori* clear that a similar degree of conceptual information is available to LSA in the absence of a morphologically modified (e.g. via stemming) training corpus.

6.3.1.6 Statistical Web-based Measures

Statistical Web-based Measures operate on the basis of arguably the largest text collection available anywhere — the web — via simple web counts using web-search engines such as Google or Altavista. They are thus not only computationally much less costly, but also circumvent the sparse data problem that vector-based methods operating on a training corpus face.[13] However, they are doing so at the expense of gathering data that tends to considerably more noisy (see Keller and Lapata (2003) for an overview of potential sources of noise in web counts).

Two methods in particular are noteworthy in this context, as they (a) are superficially similar but based on different underlying assumptions, and (b) have both been evaluated via TOEFL-style synonymy tests: the PMI-IR algorithm and LC-IR.

PMI-IR is based on Pointwise Mutual Information (PMI) to analyze statistical data collected by Information Retrieval (IR). For two words w_1 and w_2, PMI (defined as

[13] Keller and Lapata (2003) have demonstrated that web counts (at least for English) are generally useful for approximating sparse or unseen corpus data: they showed that the web can be used to obtain frequencies for unseen corpus bigrams that correlate highly with corpus frequencies, and reliably with human plausibility judgments.

$\frac{P(w_1 \& w_2)}{P(w_1)*P(w_2)}$) is proportional to a measure based on the expected counts of words and word pairs in a corpus ($\frac{Count(w_1 \& w_2)}{Count(w_1)*Count(w_2)}$), which can be estimated via web search statistics - the IR component - obtained by e.g. $\frac{Hits(w_1\ NEAR\ w_2)}{Hits(w_1)*Hits(w_2)}$.

Turney (2001) has used PMI-IR using the AltaVista NEAR operator, which searches for words within a 10-word window, and reported a performance on the TOEFL synonymy test higher than that of LSA (73%[14] compared to LSA's 64% correct answers). As has been noted by Higgins (2005), the intuitive basis for PMI-IR is quite different compared to vector-based approaches such as LSA: rather than assuming that similar words tend to occur in similar contexts, i.e. tend to have the same neighboring content words, PMI-IR assumes that similar words occur near each other.

Higgins (2005) has proposed the LC-IR (local context-information retrieval) measure, which on the surface is very similar to Turney's PMI-IR, but differs from that measure in that the NEAR operator is replaced by NEXT-TO, which requires strict adjacency of two words rather than mere proximity. In order to cancel out collocation effects, only the less frequent of the two bigrams (w_1w_2), (w_2w_1) are included in the similarity score, which is defined as

$$Similarity_{LC-IR}(w_1, w_2) = \frac{min(Hits(\text{“}w_1w_2\text{”}), Hits(\text{“}w_2w_1\text{”}))}{Hits(w_1)*Hits(w_2)}$$

Higgins claims that the additional requirement of strict adjacency is of crucial importance to LC-IR, as it practically ensures the implementation of the above-mentioned 'parallelism' assumption. The reason given by Higgins is that since search engines such as Altavista and Google ignore punctuation marks (such as commas or slashes) even if the search term is quoted, LC-IR assigns high scores to word pairs which often occur as conjoined items or other equative, or implied conjunctive, contexts. This, according to Higgins, enables the isolation of word pairs which exhibit a high degree of grammatical parallelism "because the equative uses [...] virtually guarantee parallel use of the terms" (Higgins, 2005, p. 12). However, the caveat should be noted that the removal of punctuation marks does not only give additional weight to conjoined contexts, it can also lead to the inclusion of noise such as false positives in the case of e.g. phrase boundaries.

While cautioning that LC-IR "is by no means an ideal implementation of the parallelism assumption", Higgins hypothesized that the clearer basis in parallelism (in

[14] 74% if the NEAR operator is augmented with 'and NOT', which tends to reduce the equal weight-effect for antonyms in relation to synonyms. At the time of writing, the NEAR operator is no longer supported by Altavista.

comparison to PMI-IR) should lead to a higher performance in word similarity rating tasks. To investigate this hypothesis, he compared the performance of LC-IR to other similarity measures (namely LSA, Random Indexing, PMI-IR, and a Thesaurus-based measure using Roget's Thesaurus (Jarmasz and Szpakowicz, 2003)) on the task of correctly answering multiple-choice synonym test items (the synonym section of the TOEFL test, 50 ESL questions used by Turney (2001), and a set of 300 items from the Reader's Digest Word Power (RDWP) feature). Higgins found that LC-IR outperformed all other measures on the TOEFL and RDWP test, as well as on the overall average [15] (for TOEFL, LC-IR achieved a score of 81.3% compared to LSA's 64.4%, PMI-IR's 80% and Roget Thesaurus's 78.8%; the overall average scores were 76.4% for LC-IR, 76.0% for Roget's Thesaurus, and 73.0% for PMI-IR, the baseline being 25% for all test sets).

Given LC-IR's impressive performance on these tests, taken together with the fact that web counts are computationally much less costly than using a vector-space method such as LSA, LC-IR was selected as the second alternative to the 'gold standard' dictionary-based approach described in section 6.3.3.

6.3.2 Data Collection with LSA and LC-IR

For LSA, 3 training corpus versions were selected that appeared promising as test candidates due to their high correlations achieved in the sentence similarity rating task, while at the same time representing both ends of the document size range (article and paragraph-sized documents — see also chapter 5):

- Version 1: Corpus-B, 67 MB (documents are articles; full sanitizing);
- Version 2: Corpus-C, 129 MB (documents are articles; 90% sanitizing);
- Version 3: Corpus-E, 127 MB[16] (documents are paragraphs; full sanitizing).

For each of the selected training corpus versions, the dimensionality of LSA was varied from 200 to 400 dimensions (450 dimensions for version 1), in steps of 50 dimensions.

The LC-IR procedure was implemented using the Google search engine.[17] Google was preferred over Altavista, which was used by Higgins (2005), since (at the time of

[15] LSA was only outperformed on the TOEFL test, as no corresponding test results were available for the other data sets.

[16] The corpus sizes for versions 2 and 3 differ slightly from the corresponding versions in chapter 5 because different sections of the corpus had been used in each case.

[17] The Google search was performed in December 2005.

analysis) Altavista did not allow strict adjacency search using quoted search terms.[18] The search was restricted to German-language web pages to eliminate the risk of false positives caused by crosslinguistic homonyms. Even though (as noted by Keller and Lapata (2003)) the risk of two such homonyms constituting a valid bigram in another language is fairly small, is it is arguably slightly higher for German than it is for English, due to German being a considerably less common language on the Web (Grefenstette and Nioche, 2000).

Since the training corpus for LSA contained 'raw text', i.e. word forms without any stemming or lemmatization, the search terms for LC-IR were also straight word forms to ensure comparatibility, i.e. the inflectional morphology of the words was not taken into account by either method.

In order to compare the different versions of LSA among each other on the task of word similarity judgments, as well as the LC-IR method to the LSA versions, two tests were conducted: the first test consisted of a multiple-choice lexical relation test similar to the synonym test section of the TOEFL test; for the second test, a correlation analysis was conducted comparing LSA and LC-IR scores with human similarity judgments of German noun pairs. Both analyses are presented below.

6.3.3 Data Analysis I: Multiple-Choice Lexical Relation Test

6.3.3.1 Selection of Materials

The multiple-choice lexical relation test developed for the comparative analysis of the LSA and LC-IR measures was modeled after the TOEFL synonym test section, which consists of 80 synonym test questions. However, the test developed for this analysis increases the level of difficulty by extending the TOEFL synonym test in two ways: (a) it covers the entire range of lexical relations selected for the analysis — in addition to synonyms, it also includes antonyms and hyponyms/hypernyms; (b) 100 test items instead of 80 are included; (c) for each test item, 9 different choices are offered, leading to a lower baseline of 11% (compared to TOEFL's 25%). Also, for certain test items, 'misleading' detractor items had been inserted where this seemed appropriate (max. 2 such words per test item). These misleading detractors were selected in such a way as to in part resemble the target word morphologically, while at the same time being

[18] In Altavista, the search term "doctor nurse" also returns pages that contain the term "doctor and nurse", i.e. not only strictly adjacent occurrences of the two words.

clearly semantically distinct.[19]

The 100 test items consisting of the target word, its lexical relation, and 8 detractors (including 0-2 misleading detractors), were selected in the following way: The 100 target words were selected semi-randomly out of the 243 target words of the teacher data set (see chapter 3): the selection strived for an approximate preservation of the part-of-speech ratio of the teacher data, by selecting 50 target words among nouns, 25 words among verbs, and the 25 remaining words among adjectives. Within these allocations, the target words were selected at random. An overview of the classification of the test items along the two dimensions (Part-of-speech and type of lexical relation) is provided by table 6.1.

Table 6.1: Classification of Test Items for Multiple-choice Lexical Relation Test

	Synonyms	**Hypernyms**	**Hyponyms**	**Antonyms**	**Total**
Nouns	29	13	3	5	**50**
Verbs	12	7	4	2	**25**
Adjectives	9	4	1	11	**25**
TOTAL	**50**	**24**	**8**	**18**	**100**

For each of the 100 target words, one semantically related word was selected such that each of the chosen lexical relation categories was represented, with synonyms given the highest representation: the 100 chosen semantically related words clustered into 50% synonyms, 34% hyper/hyponyms and 16% antonyms.

The lexical relations were validated in the following way: for all of the lexical relation categories, a given candidate word was a confirmed as a lexical relation (either synonym, hypernym, hyperonym, or antonym) if it was either listed as such in one of the following current lexicographic resources listed in section 5.4,[20] or one of the following morphological criteria was met:

- A *Hypernym* of the target word is confirmed as such if the target word is a semantically transparent compound that can be segmented into two constituent parts (A-B), with B being the candidate hyponym to the target word (the same applies

[19]For example, for the target word *Überschwemmung* (flood) with the lexical relation (hypernym) *Naturkatastrophe* (natural disaster), the two misleading detractors *Überschneidung* (overlap) and *Überschwang* (exuberance) were included.

[20]For GermaNet, only direct-neighbor hyponym/hypernym relations (i.e. up to a path length of 1) were considered.

vice versa for hyponyms). Example: *Halle* (hall) is a hypernym of *Lagerhalle* (storage hall).

- An *Antonym* of the target word is confirmed as such if the candidate antonym is flagged by one of the opposition-indicating prefixes (e.g. *un-, in-*). Example: *lösbar* (soluble) is an antonym of *unlösbar* (insoluble).

The 'misleading' detractors discussed above were chosen where possible and the remaining slots filled with randomly chosen detractor items from the appropriate lexical category. Since all tested LSA versions operate on word forms rather than stems or lemmas, the selected word form of the target word was chosen arbitrarily; all corresponding words within the same item (lexical relation and detractors) were then selected with the same corresponding morphological inflection. The complete set of selected test items is provided in Appendix G.

6.3.3.2 Results of Analysis

Table 6.2 lists the results of the multiple-choice lexical relation test for all tested versions of LSA and LC-IR in terms of percentages of correctly predicted lexical relations, both overall and broken down into parts-of-speech and lexical relations. A lexical relation is correctly predicted if it achieved the highest cosine score (for LSA), or the highest similarity score (according to the LC-IR word similarity formula given above). The chance baseline performance for all listed percentages is 11%; the frequency baseline (most frequent word is always selected) is 28%.[21]

6.3.3.3 Discussion

The results show that LSA achieved the overall best performance in the Corpus C-129 MB version (documents as articles; 90% sanitizing), with a total score of 51% 'correct winners'; however, several other 'big corpus' versions (C and E) performed nearly on the same level. Contrary to what has been reported for the sentence similarity rating task in chapter 5, the results seem to indicate that an increase in corpus size from ca. 67 MB to ca. 125 MB does yield moderate performance benefits for the task of predicting word similarities in German; however, this finding should be regarded as preliminary at this point especially since the tasks (correlation analysis to human judgments, and performance in a multiple-choice item test) are not directly comparable.

[21] This figure is based on the word form frequencies obtained from the *IDS* (Institut für Deutsche Sprache) corpus.

Table 6.2: Percentages of correctly predicted Lexical Relations for Multiple-Choice Lexical Relation Test

Measure Tested	**LSA**															**LC-IR**
Training Corpus	Corpus B-67 MB						Corpus C-129 MB					Corpus E-127 MB				
Dimensionality	200	250	300	350	400	450	250	300	350	400	450	200	250	300	450	
Parts-of-Speech Breakdown																
Nouns (n=50)	70%	70%	68%	78%	68%	66%	66%	74%	76%	80%	76%	66%	66%	64%	60%	84%
Verbs (n=25)	16%	20%	16%	16%	12%	16%	28%	28%	28%	24%	24%	36%	36%	28%	32%	80%
Adjectives (n=25)	20%	24%	28%	20%	20%	20%	16%	16%	20%	20%	24%	32%	32%	32%	32%	68%
Lexical Relations Breakdown																
Synonyms (n=50)	44%	46%	40%	44%	36%	36%	38%	46%	50%	50%	50%	48%	50%	44%	40%	76%
Hypo/Hypernyms(n=34)	34%	34%	36%	40%	36%	36%	59%	56%	59%	62%	56%	32%	30%	30%	32%	88%
Antonyms (n=16)	10%	12%	14%	12%	12%	12%	31%	38%	31%	31%	38%	20%	20%	20%	20%	69%
Total Percentages	**44%**	**46%**	**45%**	**48%**	**42%**	**42%**	**44%**	**48%**	**50%**	**51%**	**50%**	**50%**	**50%**	**47%**	**46%**	**79%**

Compared to Landauer and Dumais's result of 64.4% accuracy on the TOEFL synonym test, the results LSA achieved for the task at hand are considerably lower; this may be either due to the increased 'difficulty' of the test mentioned above, or the morphological richness of German compared to English, leading to a decreased availability of conceptual information for LSA to draw upon (at least in a non-lemmatized or non-stemmed corpus — see also the corresponding discussion in chapter 5).

Across all LSA/corpus combinations tested here, LSA's performance for nouns is significantly higher than that for verbs and adjectives; it may be speculated that this is due to the increased number of inflectional forms for verbs and adjectives compared to nouns in German, which may entail a decrease in conceptual information available to LSA (see above). The significantly poorer performance of LSA for antonyms may be speculatively attributed to the very same phenomenon, given that ca. 70% of the antonyms tested can be found in the lexical categories of verbs and (in particular) adjectives, both of which tend to contain considerably more inflected forms per lemma than do nouns (see table 6.1).

Turning to LC-IR, the results of the multiple-choice test confirm (for German) Higgins's findings both in regard to the comparison to other vector-based similarity measures, and the performance level of LC-IR for the multiple-choice lexical item test. LC-IR outperforms LSA overall as well as in all part-of-speech and lexical relation categories; the gap is particularly striking for verbs, where LSA's performance drops considerably compared to nouns, while LC-IR performs on nearly the same level. Even for antonyms, where LC-IR achieves the comparatively lowest results, LC-IR still clearly outperforms all versions of LSA. The overall accuracy of LC-IR (79%) reported in table 6.2 is on the same level as the results found by Higgins (81% for TOEFL, 76% overall accuracy score across all synonym tests).

In the second test reported below, LSA and LC-IR were compared to human word similarity ratings in German. The versions of the training corpora used for this analysis were Version 1 and Version 2 (Version 3 was omitted as it performed slightly worse than Version 2 in the multiple-choice lexical relation test, and featured a training corpus size very similar to Version 2).

6.3.4 Data Analysis II: Correlation with Human Ratings of Noun Pair Similarities

As a second independent criterion measure to validate the LSA and LC-IR similarity ratings, human word similarity judgments have been chosen. Previous research has shown that "people can reliably judge the degree of semantic similarity between words" (McDonald, 1997), and that these judgments are remarkably consistent over time (Rubenstein and Goodenough, 1965; Miller and Charles, 1991).

6.3.4.1 Materials

The data set used for the analysis was a subset (57 out of 65 items) of Rubenstein and Goodenough's data set of 65 English noun pairs translated into German. The set[22] had been used as the basis for the comparative analysis of various taxonomy- and information content-based word similarity measures (Gurevych and Niederlich, 2005). It should be noted that the data — being a translated version of Rubenstein and Goodenough's noun pairs — are liable to the same inherent limitations as that data set, namely the small size, and the restriction to nouns. The complete set of test items, together with the corresponding human ratings, can be found in Appendix H.

6.3.4.2 Results of Analysis

The results of the correlation analysis are based on only 56 out of the 57 items total for LSA, since one word pair involved a noun not represented in any of the LSA word spaces.[23] The results of the correlation analysis are provided by table 6.3; all LSA-to-human correlations listed in the table are in the moderate range (from $r = 0.41$ to $r = 0.51$) and significant at $p < .01$; by contrast, the LC-IR-to-human correlation is low ($r = 0.24$) and below significance level.

6.3.4.3 Discussion and Conclusion

Compared to other studies comparing taxonomy- and information content-based measures of word similarity measures with human similarity judgments, the correlations found in this analysis are significantly lower; for example, Budanitsky (1999) found correlation coefficients between $r = 0.74$ and $r = 0.85$ for Rubenstein and Goodenough's (1965) and Miller and Charles's (1991) English noun pair sets, while Gurevych

[22] Unpublished, made available to the author courtesy of Gurevych, personal communication.

[23] This word pair is marked with an asterisk in Appendix H.

Table 6.3: LSA and LC-IR correlations to Human Word Similarity Ratings

Measure Used	**Corr.**	**Measure Used**	**Corr.**
LSA-Corpus-B, 67 MB, 200 D	0.41	LSA-Corpus-C, 129 MB, 250 D	0.44
LSA-Corpus-B, 67 MB, 250 D	0.42	LSA-Corpus-C, 129 MB, 300 D	0.47
LSA-Corpus-B, 67 MB, 300 D	0.45	LSA-Corpus-C, 129 MB, 350 D	0.47
LSA-Corpus-B, 67 MB, 350 D	0.46	LSA-Corpus-C, 129 MB, 400 D	0.48
LSA-Corpus-B, 67 MB, 400 D	0.49	LSA-Corpus-C, 129 MB, 450 D	0.48
LSA-Corpus-B, 67 MB, 450 D	0.51		
LC-IR	0.24		

and Niederlich (2005) reported correlations slightly above $r = 0.70$ for the German noun pair ratings.

Surprisingly, the LSA and LC-IR figures of correlation to human judgments reported above are in direct contrast to the results found for the multiple-choice lexical relation test, in that LC-IR ($r = 0.24$) is clearly outperformed by all LSA/corpus versions tested (max. $r = 0.51$). It is not clear why this is the case, especially considering that the human ratings were restricted to noun pairs, an area where LC-IR performed most strongly both in comparison to LSA, and to other parts-of-speech (even though the gap to LSA is considerably wider for other parts-of-speech). It may be tentatively speculated that the restrictedness of the data set used for the human ratings accounts for the difference in outcome at least to some extent (even though 57 noun pairs are contained in the data set, it only comprises 46 distinct nouns).

Comparing the different LSA spaces, the fact that the smaller corpus version (67 MB) fares marginally better in this test than the bigger 129 MB corpus is also slightly surprising, given that the findings in the multiple-choice test are reversed in this regard. However, given the differences in r are fairly small, this is likely a non-significant tendency. Also, it is possible that higher-dimensional versions of the 129 MB corpus (which exceeded available computational capacity) may have achieved better performance.

In sum, the results presented in this section do not conclusively point to the superiority of either LSA or LC-IR for the task of German word similarity ratings. What is more, the performance levels found for LSA and LC-IR in both studies (with the exception of LC-IR's high accuracy rate for the multiple-choice test) are generally only

on a moderate level at best, and thus less-than-optimal for the task at hand. Given these findings, the straightforward dictionary-based 'gold standard' described above (see section 6.3.3) has been selected as the measure to be used for the analysis of word similarity in the teacher data. While this approach relegates the issue investigated above (i.e. how the issue of word similarity detection can be automated for any future implementation of the model to be developed in chapter 7) to the domain of future work, it does serve the primary goal of this section: to investigate whether teachers use significantly more words in their examples that are semantically related to the target word, compared to the original sentences. The choice of a conservative, dictionary-based standard is suboptimal also because many semantic relations between words in a text are non-classical in nature[24] (as has been pointed out by Morris and Hirst (2004)) and are therefore not covered by any such approach; however, the chosen measure has the advantage of ensuring maximum precision (if low recall), and thus constitutes a 'lower-bound' estimate of the actual amount of semantic relatedness found in the teachers' word choices.

6.3.5 Analysis of Word Similarity in the Teacher Data

As has been discussed in the preceding section, a dictionary-based manual count of words in the teachers' examples (ES) that are semantically related to the target words has been chosen as the measure to assess whether teachers use significantly more semantically related words in their examples than are present in the corresponding original sentences (OS). In order to investigate this issue, a paired t-test has been conducted on the teacher questionnaire data. The t-test reveals that, on average, the ES contain significantly more words that are semantically related to the respective target word than do the respective OS (OS: M=0.2, SE=0.08; ES: M=0.8, SE=0.19). This constitutes a significant increase in the number of words semantically related to the target word from OS to ES ($t(242) = -3.20$, $p < .01$).

In sum, the use of words semantically related to the target word has been found to be a significant factor in the selection of the teacher examples, and will therefore be included in the regression model of the teacher data (see chapter 7) as one of the predictor variables.

[24] i.e. are not covered by the traditional lexical relations such as synonymy, antonymy, hyponomy.

6.4 Analysis of Significant Co-occurrences

As has been mentioned in section 6.1, lexical choices in the teachers' ES that are *syntagmatically* related to the target word have been investigated via *significant co-occurrences* of the target word. These are words that, on the basis of a statistical corpus analysis, co-occur significantly more frequently with the target word in question than would be expected by chance.

6.4.1 Materials

For the co-occurrence analysis of the teacher data, the co-occurrence analysis component of COSMAS II[25], a corpus research and analysis system operating on the *IDS* corpus, has been used.

The IDS corpus is, to the knowledge of the author, the largest currently available collection of electronic corpora of written contemporary German, containing ca. 2 billion words (see `http://www.ids-mannheim.de/kl/projekte/korpora/`). It is based to a large extent (but not exclusively) on newspaper texts and, in addition, contains a variety of literary, scientific and 'popular science' texts.

The *Cosmas II* co-occurrence analysis yields, for a given target word form or lemma, a list of significant co-occurrences of the target word; for each significant co-occurrence, the log-likelihood ratio (LLR) is reported, which can be interpreted as a measure of how much more likely the co-occurrence of two words is than their base rate of occurrence would suggest. The details of the co-occurrence analysis procedure are described below.

6.4.2 Procedure

The co-occurrence analysis has been conducted along two dimensions: for both *content* and *function* words, and both as a simple difference count of significant co-occurrences in the ES and OS, and on the basis of the difference of accumulated log-likelihood ratios.

Cosmas II's co-occurrence analysis tool offers the option of excluding function words from the analysis. For the analysis of significantly co-occurring *content* words, this option was activated in order to exclude the influence of syntactic phenomena on the analysis of semantically interesting relations; for the analysis of *function* words, the

[25] Available online at `http://www.ids-mannheim.de/cosmas2/`

option was de-activated, with only the function words included in the co-occurrence count.

Given that the *Cosmas II* analysis provides a list of significant co-occurrences of the search terms ranked according to their LLR values, the main question of interest in this section, namely whether the use of words that are significant co-occurrences with the target word plays a more significant role in the ES compared to the OS, can be recast in terms of the following two research questions:

1. Do teachers use *significantly more* significant co-occurrences of the target word in their examples compared to the respective OS?

2. Is the *degree* to which words in the teachers' examples significantly co-occur with the target word significantly higher in the ES compared to the OS?

In other words, question (1) focusses on the mere difference count of significant co-occurrences without taking into account different weights of the relative degree of co-occurrence, while question (2) is based on the relative weights of co-occurrences (as measured by the difference of the accumulated LLR ratios of OS and ES), while neglecting the absolute count of significantly co-occurring words.

The following list describes the remaining analysis parameters in more detail.

- *Context window*: Three different window sizes have been used for the analysis: the default window size of ± 5 commonly used in co-occurrence statistics, as well as the the direct left and right neighbors of the target word (i.e. window sizes of +1 and -1, respectively);

- *Lemmatization*: *Cosmas II* offers this option for both the search term (target word) and the co-occurring words; all four permutations have been considered for the analysis;

- *Rules for Analysis*: All 12 permutations of the context window and lemmatization parameters have been considered for the analysis. For the simple co-occurrence count, a word is counted *once* if it appears in at least one of the 12 corresponding significant co-occurrence lists; for the accumulative LLR count, the (arbitrarily chosen) order of precedence for the choice of LLR value is: small context window size (left/right neighbors preferred over ± 5 window size) $>$ lemmatization for both target word and co-occurring words $>$ lemmatization for target word only $>$ lemmatization for co-occurring words only.

For the remaining secondary analysis parameters[26], the default values have been selected.

6.4.3 Results of Analysis

For all four combinations of content and function word co-occurrence analyses, using either a simple co-occurrence difference count or a difference count of the accumulated LLR values, paired t-tests have been conducted. The results are presented in table 6.4.

Table 6.4: Significance Analysis Results for Significant Co-occurrences

Word Type	**Content Words**				
	Mean OS	**St.Err.**	**Mean ES**	**St.Err.**	***t(242)***
Co-occurrence Count	0.65	.049	0.89	.055	3.92**
Accumulated LLR Count	481.6	158.0	693.1	136.1	1.44
	Function Words				
	Mean OS	**St.Err.**	**Mean ES**	**St.Err.**	***t(242)***
Co-occurrence Count	0.38	.035	0.38	.036	0.00
Accumulated LLR Count	1069.5	325.3	1243.0	462.8	0.61

6.4.4 Discussion and Summary

With respect to the two research questions outlined above, table 6.4 shows that for *content* words, teachers use *significantly more* significant co-occurrences of the target word in their examples compared to the respective OS; however, the *degree* to which content words in the ES significantly co-occur with the target word (as measured by the accumulated LLR) is higher in the ES compared to the OS, but not significantly so. This result is slightly surprising, as the LLR measure constitutes a much more fine-grained instrument than a simple count of co-occurring words; however, it may be explained by the very high figures of standard error presumably caused by very high fluctuations of the possible values the accumulated LLR can take on (from 0 to ca. 25,000).

[26] e.g. granularity, autofocus; for a detailed description of these see `http://www.ids-mannheim.de/kl/misc/tutorial.html`.

For the domain of *function* words, the teachers' examples contained exactly the same amount of significant co-occurrences of the target word than the original sentences. The figure of the accumulated LLRs was slightly higher in the ES compared to the OS; however, just as was the case for content words, the difference is below significance level. These findings may be speculatively attributed to the fact that teachers were explicitly told in the instructions (see chapter 3) that the criterion for the helpfulness of their examples should relate to illustrating the *meaning* of the target word (i.e. not its usage, as would arguably be the primary function of function words).

In sum, the use of content words that are significant co-occurrences of the target word has been found to be a significant factor in the selection of the teacher examples, and will therefore be included in the regression model of the teacher data (see chapter 7) as one of the predictor variables.

6.5 Morphological Analyses

6.5.1 Introduction

This section is concerned with an investigation of specific choices of the *morphological form* of the target word that may have been selected by the teacher for a quite diverse range of reasons that may be summarized as simplification or illustration purposes (depending mostly on the respective part-of-speech). In the following sections, each of the possible morphological choices hinted at by the the teacher explanations in chapter 3 are analyzed in turn, with exception of the too-general category "use specific person", and "use simpler verb form" (to the extent the latter category can be operationalized, it is arguably subsumed by the categories "use regular verb forms (for irregular verbs)" and "use most frequently used tense", both of which are analyzed under the heading "use of irregular verb forms" in section 6.5.5).

6.5.2 Frequency of Target Word Forms

It might be hypothesized that teachers use the target word in a frequently occurring inflectional form for either (a) simplification purposes, or (b) to show the form to learners in which the target word is most frequently used. The corresponding questions can be stated in the following form:

1. Is there a significant increase in the frequency of the target word forms in the ES

compared to the OS forms?

2. Are there significantly more top ranks among the ES target word form frequencies compared to the OS target word form frequencies?

6.5.2.1 Procedure

In order to investigate the above hypotheses, information on the target word form frequencies in the teacher data has been collected from the online *Wortschatz Universität Leipzig* electronic dictionary[27], which provides frequency information for German word forms in the form of log-based frequency classes ranging from $n = 0$ to ca. 24 (a frequency class of n for a word form x means that the most frequent word form in German, the article *der*, is ca. 2^n times more frequent than x).

In order to investigate hypothesis (1), a paired t-test has been conducted on the respective frequency classes of the OS and ES target word forms.[28] In order to test hypothesis (2), all word forms of the target word lemmas have been ranked according to their frequency classes, and the Pearson chi-square (χ^2) test has been conducted to analyze the top ranks *vs* non-top ranks transitions from OS to ES.

For the analysis, all 236 inflectable target words (out of the total 243 target words) have been selected.

6.5.2.2 Results of Analysis

The paired t-test revealed that, on average, the target word forms in the teachers' examples were slightly more frequent than those in the original sentences (for OS: M=14.97, SE=0.187; for ES: M=14.88, SE=0.187). However, this increase in frequency does not reach significance level. This result has been confirmed by the Wilcoxon signed-ranks test ($z = -1.14$, $p = 0.25$).

Table 6.5 shows that the amount of top-ranked target word forms is slightly higher in the teachers' examples compared to the original sentences (166 *vs* 158, respectively).

In order to find out whether this difference of top-ranked frequencies is significant, the Pearson chi-square (χ^2) test has been conducted. The χ^2 test revealed that the

[27] Available at `http://wortschatz.informatik.uni-leipzig.de/index_js.html`

[28] The use of the parametric t-test is problematic in this case, as the assumption of normally distributed data is usually violated in the case of word frequencies which are roughly distributed according to Zipf's law (Zipf, 1949). Also, the t-test's assumption of homogeneity of variance is violated. Therefore, the non-parametric Wilcoxon signed-ranks test has been conducted as well.

Table 6.5: Top Ranks/Other Ranks Classifications in Teacher Data Target Words

	OS	**ES**	**TOTAL**
Top Ranks	158	166	**324**
Other Ranks	78	70	**148**
TOTAL	**236**	**236**	**472**

difference in top-ranks *vs* other-ranks classifications is non-significant (χ^2=0.63, $p = 0.43$).

6.5.2.3 Discussion

As far as both hypotheses outlined above, the paired t-test and χ^2-test analyses have confirmed the expected increase in frequency of the target word form, but have also shown that it remains well below significance level, both in terms of the average increase in frequency, and the number of top-ranked frequencies associated with the target word forms. Therefore, it has to be concluded that frequency of the target word form does not constitute a significant criterion that teachers employ in their selection of the morphological form of the target word.

6.5.3 Nouns: Indication of Noun Gender

In case the target noun had been used together with an article form which is either entirely non-indicative with respect to the gender of the noun (e.g. plural article forms such as *die* (definite) or *einige*, *ein paar* (indefinite)), or only partially indicative (e.g. indefinite article singular *ein*, or conflated preposition-articles *im, vom* which suggest either masculine or neutrum), the teacher may have selected a singular form of the article (masc. *der/ein*, fem. *die/eine*) in order to unambiguously indicate the gender of the unknown or difficult target noun.

In order to investigate this hypothesis, the OS to ES "gender indication" transitions were inspected for all target nouns the genders of which are not unambiguously indicated in the OS. Of all 112 target nouns in the teacher data, this criterion is met by 66 OS. Out of these items, 26 (ca. 39%) pairs feature transitions to unambiguous gender indication, while 60 (ca. 61%) do not. This means that specific article usage to indicate the gender of the corresponding noun is clearly not a significant criterion that teachers

employ in their selection of the morphological form of the target word.

6.5.4 Adjectives: Indication of Adjectival Usage

In German, most adjectives can be used attributively (A), predicatively (P) and adverbially (V), but some adjectives are "defective" and cannot be used in all 3 ways. The following theoretically possible 6 permutations are realized in German in the following way:

- **APV**: non-defective: all 3 usages of the adjective possible — very common (e.g. *schön* (pretty), *billig* (cheap) etc.);
- **AP-**: no adverbial usage possible, the adjective can only describe persons and things (entities), but not actions (e.g. *neblig* (foggy), *viereckig* (rectangular)) — common;
- **A-V**: no predicative usage possible (e.g. *völlig* (completely), *wöchentlich* (weekly)) — rare, only 4 instances in teacher data;
- **-PV**: not realized;
- **A–**: only attributive usage possible (e.g. *dortig* (there), *vermeintlich* (putative), *steuerlich* (tax-related)) — rare, only 2 instances in teacher data;
- **-P-**: only predicative usage possible (*quitt* (even), *meschugge* (crazy) , *leid* (tired of)) — very rare, no instances in teacher data.

This classification, together with the fact that adverbially used adjectives in German are not marked by an adverb-indicating suffix such as English *-ly*, suggests the following hypothesis for teacher treatment of adjectival target words classified as either **APV** or **A-V**: if the word is used as either **A** or **P** in the OS, teachers will tend to use it as **V** in the ES, and vice versa. More specifically:

- If the word is used as **A** or **P** in the OS, teachers will tend to use the word as **V** in the ES to show it can be used adverbially;
- If the word is used as **V** in the OS, teachers will tend to use the word as **A** or **P** in the ES to show it that is an adjective, not an adverb.

Table 6.6: AP/V Transitions for APV/A-V Adjectives in Teacher Data Target Words

OS/ES	AP	V	TOTAL
AP	14	1	**15**
V	6	12	**18**
TOTAL	**20**	**13**	**33**

In order to investigate this hypothesis, the **AP/V** transitions were inspected for all 33 **APV/A-V** target adjectives; the transitions are summarized in table 6.6 for ES and OS.

In contrast to the hypothesis above, table 6.6 shows that the vast majority of transitions (ca. 79%) do not change from **AP** to **V** or vice versa. Thus specific adjectival usage along the lines of the hypothesis stated above is clearly not a significant criterion that teachers employ in their selection of the morphological form of the target word.

6.5.5 Irregular Verbs: Use of Regular and Irregular Verb Forms

Irregular verbs in German exhibit a stem-vowel change in certain inflected forms; it may thus be speculated that, for irregular verbs, teachers may change regular verb forms in the OS to irregular forms in the OS to show that the verb is irregular, and irregular forms in the OS to regular forms in the ES for simplification purposes.

In order to investigate this hypothesis, the regular/irregular form transitions were inspected for all 13 irregular verbs among the teacher data target verbs; the transitions are summarized in table 6.7 for ES and OS.

Table 6.7: Regular/Irregular Verb Form Classifications for Irregular Target Verbs in Teacher Data

OS/ES	Regular	Irregular	TOTAL
Regular	4	0	**4**
Irregular	5	4	**9**
TOTAL	**9**	**4**	**13**

In contrast to the hypothesis above, table 6.7 shows that the majority of transitions

(8 out of 13, ca. 62%) do not change from regular to irregular forms or vice versa. All attested transitions occur from irregular (OS) to regular (ES), suggesting that simplification, but not demonstration of the irregularity of the verb, may influence teachers' choice of the target verb form. However, given the majority of non-transitions and especially the very small size of the data set, which hardly allows any definite conclusions to be drawn, the change of regularity status for irregular verbs will not be considered further in the model of teacher criteria to be developed in chapter 7.

6.5.6 Indication of Non-separable Prefixes

Despite the exclusion of target verbs with separable prefixes from the data set (see chapter 3), the separability of prefixes may still indirectly figure into the teachers' choices of the morphological form of the (inseparable) target verb, in that the teacher may want to choose a form of the verb that shows to the student that it is not a verb with a separable prefix. With this hypothesis in mind, verb forms of German verbs *without* a separable prefix can be classified into the following 3 groups, in ascending order of "non-separability" indication strength:

1. *Non-indicating forms*: These do not indicate whether or not the verb is separable either because this form is never indicative of this aspect, or not indicative in this particular syntactic construction (example: infinitive verb forms may or may not fall under this category depending on syntactic context, e.g. main or subordinate clause);

2. *"Weak" indicators*: These are forms that *indirectly* indicate the non-separability of the verb not because a verb with a separable prefix would appear separated in this position, but because of the *lack* of morphological indicators of a separable prefix verb (such as the past participle infix *-ge-* in e.g. *auf-ge-fallen*, or the infinitival infix *-zu-* in e.g. *auf-zu-fallen*). Since these are indirect indicators that also presuppose an advanced knowledge of German grammar intricacies on the part of the learner, they can be considered weak clues only;

3. *"Strong" indicators*: These 'directly' indicate the non-separability of the verb in the sense that, in the given context, the corresponding verb form of a separable-prefix verb would appear as a separated form here.

On the basis of this classification, all OS and ES target verb forms of verbs that could be considered plausible candidates for being verbs with a separable prefix (by

virtue of containing more than 2 syllables; 30 target verbs have qualified) have been assigned categories from 0 (non-indicative) to 2 (strongly indicative). The hypothesis to be tested below is the following: is there a significant increase in the degree to which the ES target verb form indicates that the target verb has a non-separable prefix (compared to the OS target verb form)?

To address this question, both a paired t-test and a Pearson chi-square (χ^2) test have been conducted on the 30 data items.

6.5.6.1 Results of Analysis

The paired t-test revealed that, on average, the target verb forms in the teachers' examples were slightly more indicative of the non-separability of the verb than those in the original sentences (for OS: M=0.90, SE=0.121; for ES: M=0.97, SE=0.112). However, this increase in indicativeness does not reach significance level.

Table 6.8 shows the classifications of the OS and ES target verb forms with respect to the indicativeness of the non-separability of the target verb.

Table 6.8: Indicativeness of Non-separability of OS/ES Target Word Forms

	OS	**ES**	**TOTAL**
Non-indicative	8	6	**14**
Weak indicator	17	19	**36**
Strong indicator	5	5	**10**
TOTAL	**30**	**30**	**60**

In order to find out whether this difference in classification is significant, the Pearson chi-square (χ^2) test has been conducted. The χ^2 test revealed that the difference in indicativeness classifications among OS and ES is non-significant (χ^2=0.40, $p = 0.82$).

In contrast to the hypothesis above, the above analyses show that the increase in indicativeness with respect to the non-separability of the target verb is not a significant criterion that teachers employ in their selection of the morphological form of the target word.

6.5.7 Summary of Morphological Analyses

In sum, the analyses of the OS and ES target word forms in this section have shown that considerations of the morphological form of the target word in the ES have not been significant criteria in the teachers' choice of example sentences. They will therefore not be considered in the model of the teacher criteria to be developed in chapter 7.

6.6 Summary

This chapter addressed the question of whether specific lexical choices — either with regard to paradigmatic relations to the target word, significant co-occurrences of the target word, or choices relating to the morphological form of the target word itself — are significant factors in the teacher examples. The question has been answered in the affirmative for paradigmatic relations and significant co-occurrences, while morphological choices of the target word form have been found to be non-significant factors. As a side issue pertaining to word similarity measures, LSA as a representative of vector-space word similarity measures, and LC-IR as representative of statistical web-based approaches have been considered with respect to their suitablity for this study. The conclusion drawn was that their performance in German word similarity rating tasks was less than optimal, motivating a conservative, dictionary-guided 'manual' approach to identifying paradigmatic relations. While it is conceivable that this approach can be automated to a large extent insofar as many current dictionaries are available in electronic form, it is suboptimal due to the low expected recall (see discussion in section 6.3.4.3). Clearly, further research on this issue will be required before the automatization of the word similarity detection task can be considered for any future implementation of the model to be developed in chapter 7. Given that the analysis results for the larger test set — the multiple-choice lexical selection test — are quite encouraging for LC-IR, this measure should be tested on a larger and more comprehensive data set than the 57 noun pairs available for this study.

Chapter 7

Modeling the Teacher Criteria with Logistic Regression

7.1 Introduction

This chapter presents the modeling of the teacher criteria using logistic regression. It provides a description of how the input data for the logistic regression models were derived, as well as a motivation for, and a description of, the development of three different models. Two of these models, Nolex-A and Nolex-B, do not incorporate any filters to rule out difficult vocabulary; the third model, Lex4000, does. The corresponding Lexical Complexity Constraint, which motivates the Lex4000 model, is also discussed in this chapter.

The results of the logistic regression analysis are discussed for all three models; the chapter also includes a parameter evaluation of two of the models with regard to (a) the general behavior of the models in terms of their output, and (b) approximate frequency thresholds (for potential target words) associated with the models. It is this parameter evaluation that also motivates the development of the second model without lexical filters, Nolex-B, which differs from Nolex-A in that it lacks the interaction term included in that model.

We have seen in chapters 4 to 6 that three criteria turned out to be significant factors in the analysis of the teacher data (excluding syntactic complexity which is used as a pre-filter for the selection of example sentences):

- Similarity between the original and the example sentence as measured by the LSA cosine value;

- The difference between the number of significant co-occurrences in the original sentence and the example sentence;
- The difference between the number of semantically related words in the original sentence and the example sentence.

This situation lends itself to regression analysis, the predictor variables being the three significant factors above, and the outcome variable corresponding to the degree of helpfulness of the examples. Since, for the purposes of this study, helpfulness should be seen in terms of a binary dichotomy between 'helpful' and 'not helpful' (teachers have only been asked to provide *maximally helpful* examples), binary logistic regression has been chosen as the statistical method best suited to the current analysis.[1]

The basic principle of logistic regression applied to the data at hand is that the regression model predicts the probability of an example sentence being helpful based on observations of whether or not example sentences in the training set have been judged helpful or not helpful. For the current task with three predictor variables, the resulting regression model is described by the regression equation (cf. (Field, 2005, p. 220)):

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \varepsilon_i)}}$$

where Y is the event of the example sentence being helpful, $X_1 \ldots X_3$ are the three predictors described above, $b_1 \ldots b_3$ are the coefficients or weights attached to the predictors, and ε_i is a residual term.

While helpful examples are available in the form of the teacher example sentences, the lack of negative (unhelpful) example sentences means that these examples had to be selected in a post-hoc fashion based on the criteria for positive examples (see section 7.3).

The logistic regression model, applied to corpus sentences, thus provides a way of ranking the expected helpfulness of the corpus sentences (in relation to a given original sentence) on the basis of their respective $P(Y)$ values.

The current chapter is organised as follows: section 7.2 describes the Lexical Complexity Constraint underlying the Lex4000 model, while section 7.3 discusses the selection of the input data for the logistic regression models, which consist of both positive (helpful) and negative (unhelpful) examples. In the further sections, the results of

[1] A succinct summary of the advantages of logistic regression over discriminant analysis, a technique also used for distinguishing categorical data, based on a set of variables, is given in Howell (2002, p. 583).

the three logistic regression analyses are presented (section 7.4) together with testing of the models on unseen data (section 7.5). A parameter evaluation regarding the general behavior of the models and approximate frequency thresholds is discussed (section 7.6). The summary of the chapter is provided by section 7.7.

7.2 Lexical Complexity Constraint

The purpose of the Lexical Complexity Constraint (LCC) is to exclude from the selection process potential example sentences containing vocabulary which, for the students, is likely to be too difficult, or at least as difficult as the target word in need of explanation. While it is clear that any such constraint should be frequency-based, it is less obvious where exactly that threshold should be placed. Inevitably, this will to a large extent depend on the level of vocabulary mastery attained by the students using the system.

For the purposes of this study, Langenscheidt's Basic German Vocabulary (BGV) (Langenscheidt, 1991) of ca. 4,000 words was chosen as the basis for the LCC. The BGV consists of a core group of ca. 2,000 words most frequently used words, together with an 'expanded core' of the next 2,000 most frequent words. In total, the BGV accounts for ca. 85-90% of all written and oral communication in German (Langenscheidt, 1991, p. VII). Despite frequency being the main criterion for the selection of the BGV, other factors considered were familiarity and usefulness of the word in everyday conversation (Langenscheidt, 1991, p. VIII).

While less conservative, lower-frequency thresholds could reasonably be assumed for more advanced learners of L2 German, the reason for selecting a relatively conservative constraint such as the BGV was to ensure that the LCC lets a minimal number of words that present difficulties to learners "slip through the net". This seemed especially advisable considering that in the teacher study producing the example sentence data, several target words[2] were selected by teachers that are contained in the BGV, even though their respective student groups could be considered reasonably advanced learners of L2 German for whom syntax did not present a serious problem when reading German texts.

However, the BGV does not include certain groups of words that can also be assumed to be known to the target group of students with L1 English, namely numerals,

[2] e.g. *Klage* (suit, complaint) or *bislang* (so far, until now).

proper nouns and *close cognates*[3]. Therefore, an expanded version of the BGV that also included these groups of words was used as the final basis for the LCC: only sentences containing words included in the expanded BGV were considered to meet the LCC and were selected for the regression model incorporating the constraint (the Lex4000 model).

7.3 Input data for the Logistic Regression Models

A logistic regression analysis has been conducted for the two training sets Nolex and Lex4000 described below. Both the Nolex and the Lex4000 training sets are comprised of 195 positive example sentences and 195 negative examples. Both sets constitute 80% (2x N=195) of the complete data (2x N=243); the remaining 20% were retained for testing on unseen data.

7.3.1 Selection of Positive Examples

For both Nolex and Lex4000, 195 example sentences were used as positive input to the regression model. The 195 sentences were selected at random out of the 243 teacher examples. The reason for also using the teacher examples for the Lex4000 set is that it is reasonable to assume teachers avoided difficult vocabulary in their examples as much as possible (even though many teacher examples went beyond the relatively strict limitation that is the enhanced BGV set in their choice of vocabulary).

7.3.2 Selection of Negative Examples

Any attempt to develop a model based on the analysis of the available teacher data has to take into account an important limitation of the teacher study described in chapter 3: teachers were only asked to provide example sentences they considered to be the *most helpful* to the students, based on the assumption that it would be much easier for teachers to think of most helpful example sentences rather than of the most unhelpful ones. Even if teachers had been considered equally capable of providing unhelpful examples, asking them to provide the *maximally* unhelpful ones would have likely

[3]For the purpose of this study, *close cognates* are considered words that are spelt exactly or almost the same (addition, deletion or alteration of one letter at most) as their English counterparts, and are not "false friends".

produced many minimally short sentence constructions that are both highly artificial and, while very unhelpful, unlikely to yield any insight into the underlying criteria.

This means that any model based solely on the teacher data only has *positive* data (most helpful sentences) available as input. However, this constitutes a problem for the training of the model, which requires *negative* input as well, i.e. sentences that have shown to be most unhelpful to students. In order to circumvent this dilemma, the most reasonable (while certainly not optimal from a methodological point of view) solution appeared to be the post-hoc selection of negative examples based on the analysis of the *positive* teacher data available. In terms of the selection of negative examples, this meant *minimizing* each of the three significant factors for positive examples described above.

This goal was achieved in the following way: first, for each of the 50% most frequent target words, a pool of 10 candidate sentences was pre-selected such that the difference of significant co-occurrences (X_2 in the regression equation in section 7.1) and lexical relations (X_3) between the respective original and example sentence was minimized. Sentence length (see chapter 4) served as a pre-filter for sentence selection. Minimum values were achieved by only allowing example sentences with zero significant co-occurrences and lexical relations into the candidate pool. Second, for each of the candidate sets, two sentences with the lowest LSA cosines as the measure for sentence similarity (X_1) were picked as the final selections for negative examples.[4] This selection process differed for the Nolex and Lex4000 models in that for Lex4000, only sentences meeting the LCC constraint (see section 7.2) were allowed into the candidate pool.

7.4 Results of the Logistic Regression Analysis

The following subsections present an analysis of the correlations and collinearity between the three predictor variables (subsection 7.4.1), followed by a presentation of the results of the logistic regression analysis for each of the three models in the remaining subsections. The motivation for a second model without lexical complexity constraints

[4]The LSA training corpus used for this analysis is almost identical to the training corpus of choice in the semantic similarity analysis (see chapter 5), namely the FR Corpus1a (67 MB; 350 D), the only difference being the addition of the negative example sentences selected for the analysis. For the positive examples, the existing cosines from the analysis in chapter 5 were used, which differ minimally from the cosines one would obtain if the analysis were run anew. The difference affects only the 3rd and higher decimal places, and is likely due to either rounding errors caused by a switch in the operating system, or by the correction of a few typos in a small section of the training corpus.

(Nolex-B) will be discussed in section 7.6.1.

7.4.1 Correlations and Collinearity

Taken together, the correlations and collinearity analyses of the three predictor variables provide a diagnostic for the detection of multicollinearity, a potential problem for multiple regression models. Multicollinearity is caused by strong correlations between the independent predictor variables and leads to inflated variances of the parameter estimates. This in turn carries the risk of erroneous interpretations of the regression model.[5] It was therefore decided to inspect the correlations and collinearity diagnostics before running the actual regression analysis.

Table 7.1: Correlations (Pearson's r) between the 3 predictors for the Nolex / Lex4000 models

Predictor	**Sentence Similarity**		**Co-occurr. (Diff)**		**Lex. Rel. (Diff)**	
Model	**NOLEX**	**LEX4000**	**NOLEX**	**LEX4000**	**NOLEX**	**LEX4000**
Sent. Similarity	1	1	.259**	.230**	.119**	.117*
Co-occurr. (Diff)			1	1	.066	.066
Lex. Rel. (Diff)					1	1

Table 7.1 provides the results of the correlation analysis[6] for the three predictors (** indicates significance at $p = 0.01$, * significance at $p = 0.05$).[7]

Despite the significant (but low) correlations between Sentence Similarity and the other two predictors, Difference of Co-occurrence and Difference of Lexical Relations, the collinearity diagnostics of the VIF (Variance Inflation Factor) and tolerance coefficients (reported in table 7.2) indicate that multicollinearity is not a problem for the data at hand.[8] The condition indices (all fairly similar) and variance proportions (no two

[5] For instance, multicollinearity increases the risk of a Type II error with respect to the statistical significance of individual independent variables (i.e. a good predictor is found non-significant and rejected from the model (Field, 2005, p. 174)).

[6] The correlation and collinearity analyses were carried out in the complete data sets (N=486), not just the 80% training set.

[7] These results correspond to those of the non-parametric Spearman's rho (except that in Spearman, the Sentence Similarity *vs* Difference of Lexical Relations for Nolex is significant only at $p = 0.05$ (two-tailed).

[8] The statistics literature cites VIF >10, average VIF $>> 1$, tolerance $<.2$ as potential indicators of a collinearity problem (Bowerman and O'Connell, 1990; Menard, 1995).

Table 7.2: Collinearity Diagnostics I (Tolerance and VIF) for Nolex and Lex4000

Predictor	**Tolerance**		**VIF**	
	NOLEX	**LEX4000**	**NOLEX**	**LEX4000**
Sentence Similarity	.922	.937	1.084	1.068
Co-occurrences (Diff.)	.932	.945	1.073	1.058
Lexical Relations (Diff.)	.985	.985	1.016	1.015

Table 7.3: Collinearity Diagnostics II (Condition Indices and Variance Proportions) for Nolex and Lex4000

					Variance Proportions							
Dim.	**Eigenvalue**		**Condition Index**		(Constant)		Sentence Similarity		Co-occurr. (Diff.)		Lex Rel. (Diff.)	
	NOLEX	**LEX4000**	**NOL.**	**LEX4.**	**NOL.**	**LEX4.**	**NOL.**	**LEX4.**	**NOL.**	**LEX4.**	**NOL.**	**LEX4.**
1	1.587	1.610	1.00	1.00	.19	.19	.18	.17	.02	.03	.06	.05
2	1.114	1.097	1.19	1.21	.03	.02	.05	.04	.55	.54	.15	.20
3	0.900	0.894	1.33	1.34	.02	.01	.10	.10	.11	.16	.79	.75
4	0.400	0.399	1.99	2.01	.76	.78	.67	.68	.32	.28	.00	.00

predictors have high proportions on the same small eigenvalue) also suggest that multicollinearity is not a cause for concern for these data (cf. table 7.3). This conclusion is corroborated by the absence of suspiciously high standard errors of the predictor coefficients in the logistic regression analysis below.

7.4.2 The Nolex-A Model: Results of the Logistic Regression Analysis

For the logistic regression analysis of the Nolex-A model, all potential two-way and three-way interactions between the three main independent variables were entered into the model in addition to the variables themselves. The backward:LR (backward stepwise - removal criterion: likelihood ratio) method was used for this analysis due to the general agreement in the statistics literature that stepwise methods are preferable in contexts of predictive and exploratory research (as opposed to theory testing) (Menard, 1995, p. 54). The resulting model includes four significant predictors (the three main variables plus the interaction between Sentence Similarity and Difference of significant co-occurrences).[9]

7.4.2.1 Overall Fit of the Model

The overall fit of the final model (tested against the constant-only baseline model) after the fourth reduction step is significant at $\chi^2(4) = 247.4, p < .001$. This result indicates that all four predictors of this model — the three main variables plus the interaction term between Sentence Similarity and Difference of co-occurrences, see section 7.4.2.3 — as a set reliably distinguish between helpful and non-helpful example sentences. Overall, the final model accounts for 45.8-62.6% of the variance in helpfulness (depending on which measure R^2 is used)[10].

The highly significant value of the Hosmer & Lemeshow goodness-of-fit statistic ($\chi^2(8) = 40.143, p < .001$) indicates that the observed data are significantly different from the model's predicted values. However, since the sample size (N=390) is quite large, this is to be expected and does not necessarily imply a poor fit of the model.

[9]Note that in this case, the forward:LR stepwise method yields a different model that does not include the interaction term between Sentence Similarity and Co-occurrence as a significant predictor. This model also happens to be identical to the Nolex-B model based on the forward:LR method without interactions that is presented in section 7.4.3. The model based on the backward:LR method was preferred because the forward-stepwise method is more likely to produce a type II error (rejection of a predictor that is in fact significant) due to suppressor effects (Field, 2005, p. 227).

[10]Cox & Snell R^2: .470, Nagelkerke R^2: .626, Hosmer & Lemeshow R^2: .458

7.4.2.2 Classification Ability

The classification table (table 7.4) shows that the final model correctly classifies 87.2% of all cases (Recall: 82.1% of all observed positive cases; Precision: 91.4% of all expected positive cases).

Table 7.4: Classification table for the Nolex-A Model

Observed		**Predicted**		
		Helpfulness		Percentage
		0	1	Correct
Helpfulness	0	180	15	92.3
	1	35	160	82.1
Overall Percentage				**87.2**

7.4.2.3 Variables in the Regression Equation

The significance values of the Wald statistics for each predictor (table 7.5) indicate that all three variables significantly predict the helpfulness of the example sentence. Sentence Similarity and Co-occurrences are the most significant predictors (at $p < 0.01$), while Lexical Relations are only significant at $p < 0.05$. Of the four potential interactions between the variables, only Co-occurrence by Semantic Similarity is a significant predictor (at $p < 0.05$). As expected, the Exp(B) values of the three main predictors indicate that as the values for each of the variables increase, so do the odds of the example's helpfulness. As none of the intervals cross 1, all predictors can be considered reliable predictors of helpfulness.

7.4.2.4 Examination of Residuals

Residuals of the model have been inspected to isolate points for which the model fits poorly, as well as points that exert an undue influence on the model. To assess the latter, the influence statistics of Cook's distance, DFBeta (standardized Cook's), and leverage values have been examined.

- **Cook's distance/DFBeta**: Cook's distance is a measure of the overall influence of a case on the model; values greater than 1 are a possible cause for concern

Table 7.5: Variables in the Regression Equation for Nolex-A

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
							Lower	Upper
SENT. SIM.	21.718	3.296	43.405	1	.000	2.70E+09	4.23E+06	1.73E+12
COOCCUR	1.699	.249	46.542	1	.000	5.470	3.357	8.913
LEXREL	1.669	.721	5.361	1	.021	5.308	1.292	21.810
COOCCUR by SENT.SIM	-7.961	3.650	4.757	1	.029	.000	.000	.446
Constant	-0.487	.177	7.565	1	.006	.615		

(Cook and Weisberg, 1982). DFBeta is a standardized version of Cook's distance. Although 7 cases have unusually high values for Cook's (above 0.15), only one exceeds the recommended OK threshold of 1 (1.05). This case also has a DFBeta value above the OK threshold of ± 1 and is thus likely to have an undue influence on the model. Of the 6 cases with suspiciously high but still below-threshold Cook's values, 2 have above-threshold DFBetas and thus may have an undue influence on the model. None of the cases with non-suspicious Cook's values have above-threshold DFBetas.

- **Leverage values**: The expected leverage is 0.01=(k+1)/N, with k=4 (number of predictors) and N=390. The recommended threshold that should not be exceeded is 3*(k+1)/N=0.038. Of the 20 cases (ca. 5%) above the threshold, only 3 also have unusually high values for Cook's; of these, one case also has an above-threshold DFBeta.

- **Normalized residuals**: These should have values of less than ± 3 or ± 2 (depending on how strict the threshold is supposed to be). 7 cases (ca. 2%) have values exceeding the upper threshold of ± 3; 16 (ca. 4%) cases exceed ± 2. Of the former, one case also has 2 above-threshold values for the indicators of undue influence, i.e. 1 case is likely to be both a point for which the model fits poorly, and a point which exerts an undue influence on the model. If the criteria are relaxed to lower thresholds (± 2 for normalized residuals; at least one above-threshold value for indicators of undue influence), 2 cases (less than 1%) are possibly problematic on both counts.

To sum up these findings, there seems to be little cause for concern regarding either points with an undue influence on the model, or outliers for which the model fits poorly. Regarding the former, only 2 cases (<1%) have above-threshold values for at least 2 of the 3 indicators of undue influence (Cook's, DFBeta and Leverage value). This figure is well within or below what can be expected in a sample this size; the same goes for the 2% respectively 4% of cases with above-threshold standardized residuals, as 5-10% of cases with absolute values greater than ± 2 are to be expected in a sample this size.

7.4.3 The Nolex-B Model: Results of the Logistic Regression Analysis

The second logistic regression model without the LCC, Nolex-B, differs from Nolex-A (section 7.4.2) in that the backward:LR logistic regression was run with the specific *exclusion* of all possible interaction terms as potentially significant predictors of helpfulness. Thus, the final model does not include the interaction term found to be significant in Nolex-A, while the three main predictors have been found to be significant in Nolex-B as well. The motivation to include a second Nolex model without the interaction terms in the analysis is due to the spurious behavior of the Nolex-A model with respect to the predicted probablities of helpfulness for certain combinations of predictor values; this is discussed in more detail in section 7.6.1.

7.4.3.1 Overall Fit of the Model

The overall fit of the final model (tested against the constant-only baseline model) after the first and final reduction step is significant at $\chi^2(3) = 244.0, p < .001$. This result indicates that all three predictors of this model, as a set, reliably distinguish between helpful and unhelpful example sentences. Overall, the final model accounts for 45.1-62.0% of the variance in helpfulness (depending on which measure R^2 is used)[11], which is marginally lower than the corresponding figures of the Nolex-A model.

The significant value of the Hosmer & Lemeshow goodness-of-fit statistic ($\chi^2(8) = 17.879, p < .05$) indicates that the observed data are significantly different from the model's predicted values. However, since the sample size (N=390) is quite large, this is to be expected and does not necessarily imply a poor fit of the model; the lower

[11] Cox & Snell R^2: .465, Nagelkerke R^2: .620, Hosmer & Lemeshow R^2: .451

significance value (compared to Nolex-A) even suggests that the removal of the interaction term, while leading to a marginally lower amount of variance explained by the model, does not reduce the overall fit of the model.

7.4.3.2 Classification Ability

The classification table (table 7.6) shows that the final model correctly classifies 86.7% of all cases, which is marginally lower than Nolex-A's 87.2%. A comparison of the respective Precision and Recall figures of Nolex-B *vs* Nolex-A reveals that a slight gain in Precision (91.8% compared to Nolex-A's 91.4%) is outweighed by a loss in Recall (80.5% compared to Nolex-A's 82.1%).

Table 7.6: Classification table for the Nolex-B Model

Observed		**Predicted**		
		Helpfulness		Percentage
		0	1	Correct
Helpfulness	0	181	14	92.8
	1	38	157	80.5
Overall Percentage				**86.7**

7.4.3.3 Variables in the Regression Equation

As for the Nolex-A model, the significance values of the Wald statistics for each predictor (table 7.7) indicate that all three variables significantly predict the helpfulness of the example sentence. Also as in Nolex-A, Sentence Similarity and Co-occurrences are the most significant predictors (at $p < 0.01$), while Lexical Relations are only significant at $p < 0.05$. As expected, the Exp(B) values of the three main predictors indicate that as the values for each of the variables increase, so do the odds of the example's helpfulness. As none of the intervals cross 1, all predictors can be considered reliable predictors of helpfulness.

7.4.3.4 Examination of Residuals

Residuals of the model have been inspected to isolate points for which the model fits poorly, as well as points that exert an undue influence on the model. To assess the

Table 7.7: Variables in the Regression Equation for Nolex-B

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
							Lower	Upper
SENT. SIM.	23.151	3.408	46.133	1	.000	1.13E+10	1.42E+07	9.03E+12
COOCCUR	1.452	.200	52.563	1	.000	4.273	2.886	6.328
LEXREL	1.828	.735	6.177	1	.013	6.220	1.472	26.288
Constant	-0.515	.173	8.831	1	.003	0.597		

latter, the influence statistics of Cook's distance, DFBeta (standardized Cook's), and leverage values have been examined.

- **Cook's distance/DFBeta**: Although 4 cases have unusually high values for Cook's (above 0.15), no case exceeds the recommended threshold of 1 (1.05). This is a slight improvement compared to Nolex-A. None of the cases with non-suspicious Cook's values have above-threshold DFBetas. Thus, unlike in Nolex-A, Cook's distance/DFBeta do not indicate any cases that may have an undue influence on the model.

- **Leverage values**: The expected leverage is 0.01=(k+1)/N, with k=3 (number of predictors) and N=390. The recommended threshold that should not be exceeded is 3*(k+1)/N=0.031. Of the 14 cases (ca. 3%) above the threshold, only 2 also have unusually high values for Cook's.

- **Normalized residuals**: These should have values of less than ± 3 or ± 2 (depending on how strict the threshold is supposed to be). 9 cases (ca. 2%) have values exceeding the upper threshold of ± 3; 18 (ca. 5%) cases exceed ± 2. Of the former, no case also has above-threshold values for the indicators of undue influence, i.e. none of the cases are likely to be both a point for which the model fits poorly, and a point which exerts an undue influence on the model. If the criteria are relaxed to lower thresholds (± 2 for normalized residuals; at least one above-threshold value for indicators of undue influence), 3 cases (less than 1%) are possibly problematic on both counts.

To sum up these findings, there seems to be no cause for concern regarding either points with an undue influence on the model, or outliers for which the model fits poorly.

Regarding the former, no cases (<1%) have above-threshold values for at least 2 of the 3 indicators of undue influence (Cook's, DFBeta and Leverage value). This figure is well within or below what can be expected in a sample this size; the same goes for the 2% respectively 5% of cases with above-threshold standardized residuals, as 5-10% of cases with absolute values greater than ± 2 are to be expected in a sample this size. Compared to the Nolex-A Model, the removal of the interaction term for Nolex-B has led to a very slight increase in points for which the model fits poorly, while points with an undue influence are even less likely in Nolex-B than they are in the Nolex-A model.

7.4.4 The Lex4000 Model: Results of the Logistic Regression Analysis

For the logistic regression analysis of the Lex4000 model using the backward:LR stepwise method[12], all potential two-way and three-way interactions between the three main independent variables were entered into the model in addition to the variables themselves. The resulting model includes the three main variables as significant predictors; in contrast to the Nolex-A model, none of the interactions between the variables was found to be a significant predictor.

7.4.4.1 Overall Fit of the Model

The overall fit of the final model (tested against the constant-only baseline model) after the fifth reduction step is significant at $\chi^2(3) = 223.6, p < .001$. This result indicates that all three predictors of this model, as a set, reliably distinguish between helpful and unhelpful example sentences. Overall, the final model accounts for 41.4-58.2% of the variance in helpfulness (depending on which measure R^2 is used).[13].

The highly significant value of the Hosmer & Lemeshow goodness-of-fit statistic ($\chi^2(8) = 23.958, p < .01$) indicates that the observed data are significantly different from the model's predicted values. However, since the sample size (N=390) is quite large, this is to be expected and does not necessarily imply a poor fit of the model.

7.4.4.2 Classification Ability

The classification table (table 7.8) shows that the final model correctly classifies 84.1% of all cases (Recall: 79.5% of all observed positive cases; Precision: 87.6% of all

[12]The forward:LR stepwise method yields the same model.

[13]Cox & Snell R^2: .436, Nagelkerke R^2: .582, Hosmer & Lemeshow R^2: .414

expected positive cases).

Table 7.8: Classification table for the Lex4000 Model

Observed		**Predicted**		
		Helpfulness		Percentage
		0	1	Correct
Helpfulness	0	173	22	88.7
	1	40	155	79.5
Overall Percentage				**84.1**

Compared to both Nolex models, Lex4000 performs slightly worse in terms of both overall fit of the model and classification ability.

7.4.4.3 Variables in the Regression Equation

The significance values of the Wald statistics for each predictor (table 7.9) indicate that all three variables significantly predict the helpfulness of the example sentence. Sentence Similarity and Co-occurrences are the most significant predictors (at $p < 0.01$), while Lexical Relations are only significant at $p < 0.05$. As expected, the Exp(B) values of the three main predictors indicate that as the values for each of the variables increase, so do the odds of the example's helpfulness. As none of the intervals cross 1, all predictors can be considered reliable predictors of helpfulness.

Table 7.9: Variables in the Regression Equation for Lex4000

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
							Lower	Upper
SENT. SIM.	19.702	2.880	46.786	1	.000	3.60E+08	1.27E+06	1.02E+11
COOCCUR	1.470	.195	56.696	1	.000	4.347	2.966	6.373
LEXREL	1.805	.716	6.361	1	.012	6.082	1.495	24.735
Constant	-0.551	.170	10.514	1	.001	.577		

7.4.4.4 Examination of Residuals

Residuals of the model have been inspected to isolate points for which the model fits poorly, as well as points that exert an undue influence on the model. To assess the latter, the influence statistics of Cook's distance, DFBeta (standardized Cook's), and leverage values have been examined.

- **Cook's distance/DFBeta**: Although 4 cases have unusually high values for Cook's (between 0.19 and 0.41), they are still well below the recommended threshold of 1. Of these cases, only one case also has an DFBeta above the ± 1 threshold and thus may have an undue influence on the model. None of the cases with non-suspicious Cook's values have above-threshold DFBetas.

- **Leverage values**: The expected leverage is 0.01=(k+1)/N, with k=3 (number of predictors) and N=390. The recommended threshold that should not be exceeded is 3*(k+1)/N=0.031. Of the 15 cases (ca. 4%) exceeding this threshold, only 2 ($< 1\%$) also have unusually high values for Cook's. None of the cases with an above-threshold leverage value also has an above-threshold DFBeta.

- **Normalized residuals**: These should have values of less than ± 3 or ± 2 (depending on how strict the threshold is supposed to be). 8 cases (ca. 2%) have values exceeding the upper threshold of ± 3; 19 (ca. 5%) cases exceed ± 2. Of the former, no case also has 2 above-threshold values for the indicators of undue influence, i.e. none of the cases is likely to be both a point for which the model fits poorly, and a point which exerts an undue influence on the model. If the criteria are relaxed to lower thresholds (± 2 for normalized residuals; at least one above-threshold value for indicators of undue influence), 2 cases (less than 1%) are possibly problematic on both counts.

To sum up these results, there seems to be very little cause for concern regarding either points with an undue influence on the model, or outliers for which the model fits poorly. Regarding the former, none of the cases have above-threshold values for at least 2 of the 3 indicators of undue influence (Cook's, DFBeta and Leverage value). This figure is well within or below what can be expected in a sample this size; the same goes for the 2% respectively 5% of cases with above-threshold standardized residuals, as 5-10% of cases with absolute values greater than ± 2 are to be expected in a sample this size.

7.5 Testing On Unseen Data

To assess the accuracy of the model across different samples, two methods of testing on unseen data have been applied to the Nolex and Lex4000 models: one estimated from the original model via Stein's adjusted R^2, the other based on data splitting, using the 20% of the complete data not used for the training of the logistic regression models as the validation sample.

7.5.1 Adjusted R^2

The adjusted R^2 (also known as Stein's formula) is an estimate of the average cross-validation predictive power of the model. The loss of predictive power (or shrinkage) can then be estimated by the difference between the model's R^2 and the adjusted R^2, which is given by the following version of Stein's equation (see Field (2005, p. 172)):

$$adjusted\ \ R^2 = 1 - [(\frac{n-1}{n-k-1})(\frac{n-2}{n-k-2})(\frac{n+1}{n})](1-R^2),$$

where R^2 is the unadjusted value, n is the number of cases and k is the number of predictors in the model. Table 7.10 shows that, depending on which measure R^2 is used, the estimated amount of shrinkage is around 1% for all three models. This figure is small and can be taken as an indication that all models cross-validate well.

7.5.2 Data Splitting

7.5.2.1 Shrinkage

For an assessment of how well the three models perform on unseen data, the prediction equations derived from the screening samples (N=390, 80% of all cases) were applied to the validation sample (N=96, the remaining 20%). The observed groupings on the helpfulness score were then correlated to the predicted groupings. Table 7.11 shows the Pearson correlations, the corresponding cross-validation coefficients (r^2), and associated shrinkages (comparing the latter with the corresponding R^2 based on the actual outcome scores). The fact that all shrinkage values are fairly small ($< \pm 7\%$) indicate that all models generalize well to different sets of data.

7.5.2.2 Classification Ability

The classification tables 7.12, 7.13, 7.14 for the Nolex and Lex4000 models show that for the validation sample the ratios of correctly classified items are very similar to the

Table 7.10: Adjusted R^2 and Shrinkages

MODEL	R^2 measure used	R^2	adjusted R^2	shrinkage
Nolex-A	Hosmer/Lemeshow	0.458	0.445	1.28 %
	Cox & Snell	0.470	0.457	1.25 %
	Nagelkerke	0.626	0.618	0.88 %
	score-based	0.559	0.548	1.04 %
Nolex-B	Hosmer/Lemeshow	0.451	0.441	1.00 %
	Cox & Snell	0.465	0.452	1.26 %
	Nagelkerke	0.620	0.611	0.89 %
	score-based	0.546	0.535	1.07 %
Lex4000	Hosmer/Lemeshow	0.414	0.403	1.07 %
	Cox & Snell	0.436	0.423	1.33 %
	Nagelkerke	0.582	0.572	0.98 %
	score-based	0.469	0.457	1.25 %

Table 7.11: Shrinkages for Validation Sample

MODEL	Pearson Correlation	Cross-validation Coefficient	R^2 based on outcome scores	Shrinkage
Nolex-A	0.709	0.503	0.559	5.62%
Nolex-B	0.731	0.534	0.546	1.23%
Lex4000	0.733	0.538	0.469	-6.83%

figures obtained for the screening sample. For Nolex-A and Nolex-B, both the overall classification percentage and Precision are slightly lower compared to the screening sample while Recall is slightly higher. For the Lex4000 model, all three indicators are slightly improved for the validation sample. These results are in keeping with the fairly small shrinkage values for the models reported above, and can be taken as confirmation for the assessment that all three models generalize well to different data sets.

Table 7.12: NOLEX-A: Classification table for Validation *vs* Screening Sample
(Percentages for screening sample are given in brackets)

Observed	**Predicted**			Precision	Recall
	Helpfulness		Percentage		
	0	1	Correct		
Helpfulness 0	42	6	87.5% (92.3%)		
1	8	40	83.3% (82.1%)		
Overall %			**85.4% (87.2%)**	87.0% (91.4%)	83.3% (82.1%)

Table 7.13: NOLEX-B: Classification table for Validation *vs* Screening Sample
(Percentages for screening sample are given in brackets)

Observed	**Predicted**			Precision	Recall
	Helpfulness		Percentage		
	0	1	Correct		
Helpfulness 0	43	5	89.6% (92.8%)		
1	8	40	83.3% (80.5%)		
Overall %			**86.5% (86.7%)**	88.9% (91.8%)	83.3% (80.5%)

7.6 Parameter Evaluation of the Models

Before the three models described in this chapter are evaluated in an empirical study with teachers of L2 German (see chapter 8), a parameter evaluation was carried out. This evaluation sought to address the following issues that are discussed in the following subsections:

Table 7.14: LEX4000: Classification table for Validation *vs* Screening Sample
(Percentages for screening sample are given in brackets)

Observed	**Predicted**			Precision	Recall
	Helpfulness		Percentage		
	0	1	Correct		
Helpfulness 0	44	4	91.7% (88.7%)		
1	9	39	81.3 % (79.5 %)		
Overall %			**86.5% (84.1%)**	90.7% (87.6%)	81.3% (79.5%)

1. To assess the models' general behavior with respect to the influence of the predictors on the outcome variable (the probability of the example being helpful), for each of the main significant predictors, threshold values were determined beyond which the example was rated helpful ($P(Y) > 0.5$).

2. To assess the practical usefulness of the models with respect to word frequency considerations. Given that for a given target word, depending on the frequency of that word in the corpus, even large corpora can only provide a limited number of target words, the goal was to establish approximate frequency thresholds below which the models cannot be expected to provide example sentences that are rated helpful by the model ($P(Y) > 0.5$). These results are intended as rough guidelines for both target word selection in the main evaluation of the models, and any future implementation of the models.

7.6.1 General Behavior of the Models

For each of the three main predictors and the corresponding combinations of the two remaining predictors,[14] the following tables 7.15, 7.16, 7.17 show the thresholds that the predictor values need to cross in order for the example to be rated helpful.[15]

Looking at all three tables, the following observations can be made with respect to the general influence of each of the three predictors on the outcome values:

[14]Sentence Similarity cosines are varied in steps of 0.10, the ranges for differences of significant co-occurrences and lexical relations are [-4;+4] and [-2;+2], respectively.

[15]In each of these tables, for the Sentence Similarity thresholds, $\geq$*neg.* denotes a negative cosine threshold (as was mentioned in chapter 5, negative cosines should at least theoretically not appear in LSA and are therefore only given the generic marker here). A *no* denotes cases where the entire range of all possible cosine values [-1;+1] lies above the helpfulness threshold.

Table 7.15: Sentence Similarity Thresholds

(Nolex-A/Nolex-B/Lex4000)

Sentence Sim.	Lexical Relations								
Co-occurrence	-2			0			2		
	NOL.A	NOL.B	LEX4.	NOL.A	NOL.B	LEX4.	NOL.A	NOL.B	LEX4.
-4	>0.20	>0.43	>0.51	>0.14	>0.27	>0.33	>0.07	>0.12	>0.14
-2	>0.19	>0.31	>0.36	>0.10	>0.15	>0.18	>0.01	>neg.	>neg.
0	>0.18	>0.18	>0.21	>0.02	>0.02	>0.03	>neg.	>neg.	>neg.
2	>0.07	>0.06	>0.06	>neg.	>neg.	>neg.	no	>neg.	>neg.
4	<**0.29**	>neg.	>neg.	<**0.62**	>neg.	>neg.	<**0.95**	>neg.	>neg.

Table 7.16: Co-occurrence Thresholds

(Nolex-A/Nolex-B/Lex4000)

Co-occurr.	Lexical Relations								
Sentence Sim.	-2			0			2		
	NOL.A	NOL.B	LEX4.	NOL.A	NOL.B	LEX4.	NOL.A	NOL.B	LEX4.
0.00	≥3	≥3	≥3	≥1	≥1	≥1	≥-1	≥-2	≥-2
0.10	≥2	≥2	≥2	≥-1	≥1	≥0	≥-5	≥-3	≥-3
0.20	≥-4	≥0	≥1	≥-36	≥-2	≥-2	≥-67	≥-5	≥-4
0.30	<**3**	≥-1	≥-1	<**8**	≥-4	≥-3	<**13**	≥-6	≥-6
0.40	<**2**	≥-3	≥-2	<**5**	≥-6	≥-4	<**7**	≥-8	≥-7
0.50	<**3**	≥-5	≥-3	<**4**	≥-7	≥-6	<**6**	≥-10	≥-8
0.60	<**2**	≥-6	≥-5	<**4**	≥-9	≥-7	<**5**	≥-11	≥-11

Table 7.17: Lexical Relations Thresholds
(Nolex-A/Nolex-B/Lex4000)

Lexical Rel.	Co-occurrences								
Sentence Sim.	-4			0			4		
	NOL.A	NOL.B	LEX4.	NOL.A	NOL.B	LEX4.	NOL.A	NOL.B	LEX4.
0.00	≥5	≥4	≥4	≥1	≥1	≥1	≥-3	≥-2	≥-2
0.10	≥2	≥3	≥3	≥-1	≥0	≥0	≥-3	≥-4	≥-4
0.20	≥-2	≥1	≥2	≥-2	≥-2	≥-1	≥-2	≥-5	≥-5
0.30	≥-5	≥0	≥1	≥-3	≥-3	≥-2	≥-1	≥-6	≥-6
0.40	≥-8	≥-1	≥0	≥-4	≥-4	≥-4	≥-1	≥-7	≥-7
0.50	≥-11	≥-2	≥-1	≥-6	≥-6	≥-5	≥0	≥-9	≥-8
0.60	≥-14	≥-4	≥-2	≥-7	≥-7	≥-6	≥0	≥-10	≥-9

- All models appear to be extremely sensitive to the Sentence Similarity predictor; for the 'default' of zero difference in both lexical relations and significant co-occurrences, a cosine of ca. 0.02 is high enough for the example to be rated helpful. This degree of sensitivity of the models to very slight increases of the LSA cosine values is not in line with expectations, as sentence pairs only start to be perceived as similar at much higher cosines[16]. It may be speculated that the undue bias that sentence similarity value apparently exert over the model is due to the broader range of possible values for that predictor: in contrast to the other two predictors, which take on discrete integer values within a limited range (between ± 4 in practice), sentence similarity, being a continuous variable with values between 0 and 1, has a much broader range of possible values. This problem may be compounded by the way the model was derived (negatives were selected choosing the lowest available values for each of the variables).

- For both difference of significant co-occurrences and difference of lexical relations, assuming the default of no sentence similarity (cos=0.00), the minimal difference of 1 is enough to put the examples over the helpfulness threshold; if the sentences are rated similar (cos $>\sim 0.20$), even a negative difference does not prevent the example from being rated helpful. Again, this indicates an undue

[16] As was shown in chapter 5, a precision-recall analysis with respect to human similarity judgments revealed cosines above 0.35 as yielding the highest F scores.

bias of the models in favor of sentence similarity as the most influential predictor.

The finding that the model appears to be very sensitive to even slight increases in LSA cosines also means that the number of examples rated helpful by the models can be expected to be quite high. This is confirmed by the results of the word frequency threshold analysis (see table 7.18 below).

The tables also reveal that for the Nolex-A model, high values of sentence similarity and difference of significant co-occurrences yield a completely unexpected behavior of the model, namely that of a *helpful* rating for an example if the other indicator *falls below* a certain threshold. This effect is due to the interaction term between these two variables, which in Nolex-A is included in the model as a significant predictor. Since these results obviously run counter to the desired behavior of the model, a second Nolex model (Nolex-B) was derived that differs from Nolex-A by not including any interaction terms in the logistic regression analysis.

7.6.2 Word Frequency Thresholds

In order to establish approximate word frequency thresholds above which the models can be expected to provide sentences rated as helpful,[17] the following procedure was adopted: taking a frequency-ranked list of all target words, every fifth word was selected[18] in ascending order of frequency[19]. The input for the logistic regression models (Nolex-A[20] and Lex4000) consisted of the corresponding original sentences from the articles used for the first teacher study (see Chapter 3) and examples from the following corpus resources: Frankfurter Rundschau (FR) corpus 1992/93, IDS Mannheim corpus, and the Wortschatz Leipzig corpus.

Table 7.18 shows the selected target words together with their IDS frequency count (Column A). Column B shows the corresponding corpus counts *before*, Columns C (for Nolex-A) and D (for Lex4000) *after* the applications of any filters and corrections. These are: the syntax filter (see Chapter 4), the Word Sense Disambiguation filter[21] (see Chapter 6), the lexical complexity filter for the Lex4000 model (see sec-

[17] Since the Vocabulary Learning Environment incorporating the model is intended to be flexible with respect to the number of example sentences presented to the student for any given target word, the required minimum number of helpful sentences will vary accordingly.

[18] Except the bottom 6 target words that have a zero frequency count in the IDS corpus.

[19] based on the frequency count of the IDS corpus

[20] For this analysis, only Nolex-A was considered since the LCC can be expected to determine the number of available positive examples to a much greater extent than the relatively small difference in probability ratings between the two Nolex models.

[21] None of the target words used here were ambiguous so no WSD was needed.

tion 7.2), plus the removal of any duplicates, headlines, and incomplete sentences from the raw corpus data. Columns E and F show the number of positive-rated sentences for both models. Column G complements the frequency information by showing the other variables that have a bearing on the outcome, namely the number of significant co-occurrences and lexical relations in the *original sentence* for each target word (given here as the sum of both). For the data at hand, only target word #9 has an above-zero value of 3, which is also the highest value on record for all target words. As expected, this target word has a markedly lower number of positive-rated examples than what would be expected on the basis of frequency alone. Aside from the frequency threshold information, a notable result of this analysis is that the percentage of positive examples is surprisingly high: around 80% of all tested corpus sentences for Nolex-A, and still above 50% for Lex4000 (assuming the most common zero value in column G).

Table 7.18: Word Frequency Thresholds

(Nolex-A/Nolex-B/Lex4000)

A: IDS frequency count

B: Corpus counts before filters and corrections

C/D: Corpus counts after filters and corrections for Nolex-A/Lex4000

E/F: # of positive-rated sentences for Nolex-A/Lex4000

G: # of co-occurrences and lexical relations in original sentence

	Target Word	**A**	**B**	**C**	**D**	**E**	**F**	**G**
#1	Ramschbude	1	5	2	1	0	0	0
#2	Seitenaufprall-Schutzsystem	16	20	11	0	11 (100%)	0	0
#3	Operationsbesteck	57	86	69	5	55 (80%)	3 (60%)	0
#4	Zwingburg	70	122	65	3	54 (83%)	2 (67%)	0
#5	Garküche	173	327	185	17	155 (84%)	12 (71%)	0
#6	gesetzt den Fall	243			15		15 (100%)	0
#7	Minderwertigkeitsgefühl	317			29		26 (90%)	0
#8	draufgängerisch	372			26		14 (54%)	0
#9	Handlungsmöglichkeit	423			41		5 (12%)	3
#10	Ansteckungsgefahr	529			48		32 (67%)	0

Overall, the table shows that frequency thresholds (IDS counts) of 50 for Nolex-A and ca. 100 for Lex4000 seem reasonable estimates above which the models can be

expected to yield positive-rated sentences. These thresholds are very low and show that word frequency is only a constraining factor for the least frequent of target words.

It should be remembered, however, that these frequency thresholds can only serve as very rough guidelines, as the number of sentences rated as helpful will not only depend on the frequency of the target word, but also on the original sentence the target word happens to appear in, as has been noted in the discussion of the table columns above.

7.7 Summary

This chapter discussed how the criteria for helpful examples derived from the teacher data have been modeled using logistic regression analysis. The selection of the input data for the models was discussed, as was the Lexical Complexity Constraint which constitutes the basis for one of the models, Lex4000. The results of the logistic regression analysis for all three models were presented, and the general behavior of the models analysed in a parameter evaluation. This analysis revealed the spurious behavior of the Nolex-A model for certain value ranges and motivated the analysis of a second Nolex model, Nolex-B, without the interaction term included in the Nolex-A model. The analysis also exposed a weakness of all three models developed, namely the undue bias of the sentence similarity predictor on the probability ratings of expected helpfulness of the examples. Finally, a word frequency threshold analysis was carried out in order to ascertain approximate frequency thresholds for which potential target words can be expected to provide positive-rated example sentences.

Chapter 8

The Evaluation of the Models

8.1 Introduction

In chapter 7, three binary logistic regression models have been developed that rank potential example sentences from the corpus in relation to a given original sentence. This chapter presents the evaluation of these models.

The evaluation consists of two evaluation studies using teachers and students of L2 German as subjects, respectively. The main purpose of the evaluation is to assess the quality of the models by submitting their output to both experienced teachers and intermediate-to-advanced level students of L2 German for judgment of their respective helpfulness. In particular the evaluation seeks to answer the following questions:

1. How do the models' preferred examples compare with the gold standard of example sentences provided by an experienced teacher of L2 German?

2. How do the models' preferred examples compare with the examples provided by suitable dictionaries?

3. How do the models' preferred examples compare to each other?

4. Within each model, does the model provide consistent internal ordering, i.e. are the models' top-, medium- and bottom-ranked examples perceived as better, average and worse respectively by the experienced teachers of L2 German? In particular, are the models' top-ranked examples perceived as significantly more helpful than both random corpus selections (represented by the models' medium-ranked examples) and bottom-ranked examples?

8.2 Evaluation Study I (with Teachers)

8.2.1 Participants

As per the study described in chapter 3, participants were sought among experienced[1] of L2 German. 14 teachers (2 teachers from the German Department of Edinburgh University with L1 English and 12 teachers from language schools in Germany with L1 German) participated in the study. Every teacher was offered a remuneration for their participation.

8.2.2 Materials

The materials consisted of a questionnaire for the rating of potential examples for 20 different target words. Each target word was provided with its respective original sentence context, followed by 8-10 potential example sentences containing the word. Both the order of the target words and the order of example sentences for each target word were randomized for each participant, in order to avoid the possibility of an order bias in the final ratings.

8.2.2.1 Selection of Target Words

The 20 target words and their respective sentence contexts were selected in the following way: from each of the 17 texts that were used for the first teacher study, plus 3 additional on-line newspaper articles, one word was picked at random if it satisfied two conditions: (a) the word had not already been picked as a target word by a teacher in the first study[2]; (b) the word is not included in the expanded Basic German Vocabulary (BGV) described in section 7.2. The purpose of the latter constraint is to exclude words that the target students can be assumed to either know or straightforwardly guess from their L1, while at the same time preserving a broad range of words to choose from.

8.2.2.2 Selection of Example Sentences

Example sentences were selected from a pool of three corpora previously described: IDS Mannheim, Wortschatz Leipzig and the FR corpus. From these sources, a maximum of 100 sentences (for the Nolex models) and 50 sentences (for the Lex4000

[1] with at least 2 years of teaching experience

[2] Unless the teacher merely identified the word as difficult but failed to provide an example sentence for it.

model) was chosen for each target word in the following way: For the Nolex models, the sentences were selected evenly from all three corpora if possible.[3] For all three models, sentence length (see chapter 4) served as a pre-filter for sentence selection. An additional pre-filter consisted in the *same word sense criterion*, which followed the general "lowest common demoninator" approach adopted in this thesis with respect to word sense ambiguity (see chapter 6), i.e. — based on introspection — sentences were only *excluded* if they contained instances of the target word belonging to clearly distinct senses of the word. For the Lex4000 model, only sentences comprised entirely of the expanded BGV were considered. Nolex-selected sentences meeting the expanded BGV constraint were also selected for the Lex4000 set and then supplemented by selections from the corpora, with the preferred corpus alternating for each target word. For practical reasons, an additional constraint for example selection was that only sentences containing one of the three most frequent forms of the respective target word were considered.

For each of the 20 target words, the resulting pre-selection pool of Nolex and Lex4000 candidate sentences was analyzed according to the three relevant main factors for the three binary logistic regression models: sentence similarity and the difference of significant co-occurrences and lexically related words in the original and example sentence, respectively. The three resulting rankings of the candidate sets then served as the basis for the final selection for the teacher questionnaire: for each model, the top-ranked, medium-ranked[4] and bottom-ranked example sentences were selected.

For each target word, this set of examples was supplemented by a teacher and a dictionary example. The teacher examples were provided by an experienced teacher of L2 German, who had also participated in the first study, via an on-line web-based form presenting each target word together with its paragraph context.[5]

The dictionary examples were taken from standard monolingual dictionaries of L2 German (PONS, 2004; Langenscheidt, 2003). In cases where these sources did not yield any examples for the target words in question, additional dictionaries (Collins, 1991; WAHRIG, 2003; DUDEN, 2002) not specifically targeted at learners of L2 German (either bilingual or monolingual) were also considered. If several examples were

[3]For some infrequent words, the FR corpus did not provide enough examples; in these cases, additional examples were chosen from the other two corpora. In two cases, the low frequency of the target word meant that less than the maximum of 50 sentences (11 and 32 respectively) could be selected in total for the Lex4000 model.

[4]Due to the even number of sentences, the medium rank was taken to be whichever of the two middle sentences was closer to the average.

[5]The complete contexts of the full articles were also accessible to the teacher via links.

provided by the sources above, the example used for the questionnaire was selected at random.

Thus, the final selections for the teacher questionnaire comprised 11 categories of example sentences: 9 categories representing highest-, medium- and lowest-ranked examples for each of the three models (Nolex-A, Nolex-B and Lex4000), plus teacher and dictionary examples. The number of 8-10 example sentences for each target word on the teacher questionnaire is owing to the fact that dictionary examples were only available for 17 of the 20 target words, and that in some cases the three models yielded the same rankings for some candidate sentences.[6]

8.2.3 Procedure

In keeping with the instructions for the first study described in chapter 3, teachers were asked to imagine a computer-assisted reading environment where learners of L2 German faced with an unknown or difficult word could choose among several explanation options, one of which being example sentences. They were then asked to read the original sentence context for each target word and rate each of the following example sentences according to how helpful they perceived them to be to the students in illustrating the meaning of the word as it appeared in the original context. Each of the example sentences was accompanied by a *helpfulness* scale with values between 1 and 9, where the value of 1 corresponded to *not helpful* and the value of 9 corresponded to *very helpful*.

Teachers were explicitly asked to prevent their ratings from being influenced by the fact that some examples might appear difficult to understand as they were presented out of context. They were also told that all example sentences contained the target word in *roughly the same sense* as it was used in the original sentence. In order to ensure consistency with the first teacher study in chapter 3 that was conducted with teachers based in Scotland, the teachers based in Germany (all of whom had at least working knowledge of English) were asked to imagine that the L1 of the students

[6]The first two questionnaires used in the study differ slightly from the remaining ones in that for one of the target words, six of the model categories were represented by incorrect selections. This mistake was discovered after the first two questionnaires had been completed and was rectified for the remaining questionnaires. The error was due to an incorrect word in the original sentence, which had a slight effect on the rankings for the affected target words via the semantic similarity scores (as well as a negligibly small effect on the LSA scores for the remaining target words). The corrected rankings are based on the re-computed semantic similarity scores for the target word in question. Due to the corrected rankings being very similar to the incorrect one, it was decided to keep the ratings for the incorrectly selected sentences in the two questionnaires, since the erroneous selections can still be viewed as representative of the levels (high, medium, and low) they were intended to represent.

in question was English. The first two examples for a target word, together with its original context, are provided as an example in figure 8.1. A full sample questionnaire (with translated instructions) is provided as Appendix I.

WORD 17: eingehend (thoroughly)

Die Aufzeichnungen belegen, dass Experten auf lokaler und Bundesebene die Bedrohungen durch den Hurrikan **eingehend** diskutiert hatten.
(The records prove that local and nationwide experts had discussed the threats of the hurricane **thoroughly***).*

Ex.1: Dabei haben die Betreuer Gelegenheit, eingehend mit den einzelnen Mädchen und Jungen zu reden, ihre Probleme kennenzulernen.
(This gave the minders the opportunity to talk thoroughly to the individual girls and boys, and to get to know their problems.)

Not helpful 1 2 3 4 5 6 7 8 9 **Very helpful**

Ex. 2: Dieses Thema sei eingehend diskutiert worden.
(This topic is said to have been discussed thoroughly.)

Not helpful 1 2 3 4 5 6 7 8 9 **Very helpful**

Ex. 3: Diese Frage wurde von Experten eingehend geprüft.
(This question was examined thoroughly by experts.)

Not helpful 1 2 3 4 5 6 7 8 9 **Very helpful**

Figure 8.1: Annotated excerpt from a questionnaire page containing the target word with its original context and the first three example sentences

8.2.4 Results

8.2.4.1 General Results

Before we proceed to analyze the questionnaire data in detail, a cursory inspection of the mean ratings across participants and across target words (see table 8.1), for target words across participants (see table 8.2), and for participants across target words (see

table 8.3), provides first indications for the analysis. Since dictionary examples were only available for 17 out of 20 target words, the analysis of this category was done both with and without penalizing the missing values. The "with penalty" version assigns a score of 1 (not helpful) to the missing examples; the "no penalty" version only takes into account the 17 target words with available examples.

Table 8.1: Mean ratings across participants and target words

Model Type	Model Level	Mean	Std. Err.
Nolex-A	top	3.93	0.15
	medium	4.52	0.15
	bottom	3.46	0.14
Nolex-B	top	4.70	0.16
	medium	4.28	0.15
	bottom	3.46	0.14
Lex4000	top	4.52	0.17
	medium	3.59	0.16
	bottom	4.07	0.16
Dictionary	with penalty	4.61	0.18
Dictionary	no penalty	5.24	0.18
Teacher		6.43	0.14

As expected, table 8.1 shows that the ratings for the teacher examples are higher than those of all other categories, including dictionary examples. The dictionary examples, however, are only rated consistently higher than the models' top-ranked examples if missing entries are not penalized; in the other case, while achieving higher ratings than the Nolex-A model, they are rated at roughly the same level as the top-ranked examples of both Nolex-B and Lex4000.

As for comparisons both within and among the three models, the Nolex-B model (without the interaction term) achieves the best overall results, both in terms of providing consistent calibration (it is the only model where top-ranked examples are rated higher than medium-ranked examples, and the latter are in turn rated higher than bottom-ranked examples), and in terms of the highest overall ratings of all models for the top-ranked sentences.

A glance at the by-target word and by-participant mean ratings and standard devia-

tions (see table 8.2 and 8.3, respectively) reveals that inter-participant consistency was considerably higher than inter-target word consistency: by-participant mean ratings only range between 3.59 and 5.69, while the by-target word mean ratings are between 2.76 and 6.34.

Table 8.2: By-target word mean ratings (in descending order)

Target Word	**Mean Rating**	**St.Err.**
Wassertemperatur *(water temperature)*	6.34	0.22
Einheit *(unity)*	5.22	0.21
Todestag *(day of death)*	5.21	0.22
Meinungsumfrage *(survey)*	5.15	0.23
allerlei *(all kinds of)*	4.94	0.18
vertreiben *(to drive out, expel)*	4.90	0.22
Gebäck *(biscuits, pastries)*	4.78	0.22
Machthaber *(ruler)*	4.64	0.21
derzeitig *(current)*	4.55	0.20
reagieren *(to react)*	4.32	0.21
Gesprächspartner *(interlocutor)*	4.21	0.21
durchaus *(quite, really)*	4.20	0.20
offenbaren *(to reveal)*	4.19	0.20
eingehend *(thoroughly)*	4.16	0.20
Falle *(trap)*	3.53	0.21
Jauche *(sullage, sewage)*	3.44	0.22
neidisch *(envious)*	3.44	0.20
enttarnen *(to expose)*	3.40	0.20
Innenstadtbereich *(city center area)*	3.12	0.20
Currywurst *(curry sausage)*	2.76	0.20

In order to analyze the teacher data in greater detail and test the hypotheses outlined above, planned comparisons were carried out on the data. The planned comparisons consisted of the paired t-test and relied on the Bonferroni-adjusted alphas ($\alpha_1 = 0.05/15 \simeq 0.0033; \alpha_2 = 0.01/15 \simeq 0.0007$) to take into account that given the above hypotheses, 15 comparisons were needed in total. The results of the planned comparisons are presented in table 8.4 for both the by-participant and by-target word analysis. The table shows the respective t-values together with their associated sig-

nificance levels: t-values significant at α_1 (two-tailed) are marked "*", and t-values significant at α_2 (two-tailed) are marked "**". T-values with the "wrong sign" (i.e. the difference between the respective mean ratings is in the opposite direction than expected) are given in brackets. The dictionary ratings with penalty are marked †, those with no penalty are marked ‡.[7]

Table 8.3: By-participant mean ratings

Teacher	Mean Rating	St.Err.
#1	4.09	0.22
#2	3.65	0.16
#3	4.39	0.14
#4	4.44	0.18
#5	3.95	0.17
#6	4.43	0.20
#7	4.04	0.17
#8	5.66	0.18
#9	4.76	0.17
#10	5.69	0.18
#11	4.48	0.20
#12	4.35	0.19
#13	3.87	0.17
#14	3.59	0.17

The paired t-test analysis reveals the teacher examples to be rated significantly higher than all of the top-ranked model selections, in both the by-participant and by-target word analysis. The only exception to this is the teacher example *vs* NolexB-top comparison where the difference did not quite reach significance level.

In contrast, for the dictionary example ratings involving the penalty for missing values, no significant difference could be found when comparing the dictionary example ratings to the ratings for the top-ranked model selections. It is only when the analysis is restricted to the target words for which dictionary examples are available that at least one significant difference could be found (between the dictionary exam-

[7]The analysis for dictionary examples is based on only 17 of the 20 target words and involved only the three comparisons pertinent to dictionaries. Therefore, the corresponding Bonferroni-adjusted alphas are ($\alpha_1 = 0.05/3 \simeq 0.0167; \alpha_2 = 0.01/3 \simeq 0.0033$).

Table 8.4: t-values of planned comparisons

Category A	**Category B**	**By-Participant**	**By-Target Word**
Teacher	NolexA-top	-12.72**	-4.92**
Teacher	NolexB-top	-9.55**	-3.20
Teacher	Lex4000-top	-8.76**	-3.49*
Dictionary†	NolexA-top	-2.45	-1.24
Dictionary†	NolexB-top	(0.37)	(0.19)
Dictionary†	Lex4000-top	-0.55	-0.18
Dictionary‡	NolexA-top	-3.27*	-1.99
Dictionary‡	NolexB-top	-1.06	-0.62
Dictionary‡	Lex4000-top	-2.22	-1.03
NolexA-top	NolexA-medium	(-4.12)*	(-1.05)
NolexA-top	NolexA-bottom	3.93*	0.81
NolexB-top	NolexB-medium	2.56	0.90
NolexB-top	NolexB-bottom	9.54**	2.26
Lex4000-top	Lex4000-medium	8.11**	2.36
Lex4000-top	Lex4000-bottom	2.91	0.95
NolexA-top	NolexB-top	-6.74**	-2.47
NolexA-top	Lex4000-top	-3.49	-1.37
NolexB-top	Lex4000-top	1.21	0.45

ples and the NolexA-Top examples in the by-participant analysis). However, even for the non-penalized dictionary examples, none of the other comparisons reached significance level either.

Intra-model comparisons between the top-level and the medium- and bottom-level within a model reveal a mixed picture. For the Nolex-A model, the top-ranked sentences are rated significantly higher than the bottom-ranked sentences in the by-participant analysis; however, also in the by-participant analysis, they are rated significantly *lower* than the medium-ranked sentences. The by-target word analysis did not reveal any significant differences between the top-ranked sentences and either the medium- or bottom-ranked sentences. For the Nolex-B model, the results are more encouraging than those for Nolex-A: top-ranked examples are rated significantly higher than bottom-ranked ones in the by-participant analysis (but not in the by-target word analysis); and while the difference between top- and medium-ranked examples is not high enough to reach significance level in either analysis, at least all the t-values are in the expected directions.

Overall, the Lex-4000 model fares slightly better than the Nolex-A model but not as well as the Nolex-B model in the planned comparisons analysis. In contrast to the Nolex-A model, top-ranked sentences are rated higher than both medium- and bottom-ranked ones, with the first difference revealed as significant in the by-participant analysis. However, in contrast to the Nolex-B model, the difference between top- and bottom-ranked examples is not significant in either the by-participant or the by-target word analysis, reflecting the fact that the medium-ranked sentences are rated *lower* than the bottom-ranked ones.

Inter-model comparisons involved three comparisons between each of the top-level categories. The analysis reveals that Nolex-B top-ranked examples are rated significantly higher than Nolex-A top-ranked ones, albeit only in the by-participant analysis. None of the remaining comparisons were significant in either analysis.

A general observation with respect to the results of the planned comparisons is that significance levels are reached much more readily in the by-participant analysis. This is a surprising result, as — due to the higher number of target words ($n = 20$) compared to participants ($n = 14$) — the opposite behavior was to be expected. However, it is a result that confirms the hypothesis that the by-participant and by-target word listing of means and standard deviations above have already hinted at, namely that inter-participant consistency is notably higher than inter-target word consistency. Apparently this difference is high enough to override the difference between the number

of participants and target words.

8.2.4.2 Results by Parts-of-Speech and Frequency

Because of the apparent lack of inter-target word consistency, a more detailed analysis of smaller subgroups of target words seemed called for, in order to find out whether different subgroupings of target words made a difference for the significance levels obtained. To this end, target words were broken down according to their parts-of-speech (Nouns, Verbs, Adjectives & Adverbs) and frequency (high, medium, low)[8] categories.

Table 8.5 lists the mean ratings for target words in total compared to part-of-speech and frequency groupings, respectively. Tables 8.6 (for parts-of-speech) and 8.7 (for frequency categories) present the corresponding results of the planned comparisons via paired t-tests.[9] Since the overall picture with respect to comparisons relating to teacher and dictionary examples is sufficiently clear from the overall analysis, only the 9 remaining comparisons pertaining to inter-model and intra-model comparisons were retained for the analysis of subgroupings. A welcome side-effect of the lower number of comparisons was a reduction of the Bonferroni adjustment (which is a conservative underestimate of significance levels), so that for this analysis, the Bonferroni-adjusted alphas were ($\alpha_1 = 0.05/9 \simeq 0.0056; \alpha_2 = 0.01/9 \simeq 0.0011$). Non-penalized versions of the dictionary example ratings were not considered for this analysis.

Looking first at the mean ratings for the parts-of-speech categories, one can see that nouns clearly outperform the verb and adverb/adjective categories, as well as the total group of all target words, both in terms of absolute mean ratings for the top-ranked examples and the ordering of the three levels within a model. For nouns, the top-ranked examples of both Nolex models, as well as dictionary examples, achieve notably higher ratings compared to the average total. What is more, nouns are the only part-of-speech category for which two out of the three models exhibit the expected internal ordering of the levels; in contrast, for both verbs and adjectives, none of the models has the expected level ordering. For the Nolex-B model, which had been ordered in the expected way in the all-target word group already, the gap between the top-ranked examples and the lower-ranked ones has widened considerably (4.70 *vs* 4.28 and 3.46 in total, 5.28 *vs* 3.88 and 3.26 for nouns).

[8] high: IDS count $>$ 20000; medium: 4000 - 20000; low: $<$ 4000

[9] In both tables, t-values with the "wrong sign" (i.e. the difference between the respective mean ratings is in the opposite direction) are given in brackets.

Table 8.5: Mean ratings for parts-of-speech and frequency groupings of target words

Category	Total	Part-of-Speech			Frequency		
		Nouns	Verbs	Advj	High	Med	Low
Nolex-A-top	**3.93**	4.29	4.16	2.96	3.77	4.55	3.45
Nolex-A-medium	**4.52**	4.34	5.27	4.31	4.15	5.01	4.35
Nolex-A-bottom	**3.46**	3.26	2.79	4.46	3.79	3.56	3.09
Nolex-B-top	**4.70**	5.28	3.55	4.34	4.12	5.33	4.57
Nolex-B-medium	**4.28**	3.88	4.64	4.86	4.64	4.61	3.63
Nolex-B-bottom	**3.46**	3.26	2.79	4.46	3.79	3.56	3.09
Lex4000-top	**4.52**	4.51	4.38	4.66	4.93	5.09	3.59
Lex4000-medium	**3.59**	4.08	3.11	2.90	4.20	3.89	2.77
Lex4000-bottom	**4.07**	3.94	4.13	4.31	4.48	4.38	3.41
Dictionary	**4.61**	5.14	4.59	3.44	4.63	4.50	4.69
Teacher	**6.43**	6.42	6.82	6.14	6.46	6.60	6.23
AVERAGE	**4.33**	**4.40**	**4.20**	**4.26**	**4.45**	**4.64**	**3.90**

An inspection of the mean ratings for the frequency categories reveals a slightly less clear picture than was shown by the parts-of-speech subgroupings. High-frequency words exhibit the overall worst performance by being the only category for which no model, not even Nolex-B, has the expected internal ordering of ranking levels. Both the medium- and low-frequency groups mirror the overall results in this regard by retaining Nolex-B as the only model with the expected internal ordering. Between the two, medium-frequency words seem to fare better as they have considerably higher rating averages compared to low-frequency words.

With regard to the planned comparisons for parts-of-speech, a comparison of the t-values for nouns and other target words confirms that nouns perform considerably better than target words belonging to other parts-of-speech, and even target words in general. Considering the Nolex-A model first, while the medium-ranked sentences are still rated higher for nouns than top-ranked sentences, the difference is small and not significant in either the by-participant or by-target word analysis (in contrast to the analysis in total). For the Nolex-B model, nouns are the only group where the top-ranked examples are rated significantly higher than both medium- and bottom-ranked ones in the by-participant analysis. The Lex4000 model shows a slight improvement

Table 8.6: t-values of planned comparisons for Parts-of-Speech

		TOTAL ($n = 20$)		**Nouns** ($n = 11$)		**Verbs** ($n = 4$)		**Adj/Adv** ($n = 6$)	
Category A	**Category B**	**By-Participt.**	**By-Targetwd.**	**By-Participt.**	**By-Targetwd.**	**By-Participt.**	**By-Targetwd.**	**By-Participt.**	**By-Targetwd.**
NolexA-top	NolexA-medium	**(-4.12*)**	**(-1.05)**	(-0.32)	(-0.05)	(-3.04)	(-1.34)	(-3.62*)	(-6.72*)
NolexA-top	NolexA-bottom	**3.93***	**0.81**	5.69**	1.20	4.68**	1.77	(-5.40**)	(-2.02)
NolexB-top	NolexB-medium	**2.56**	**0.90**	6.19**	2.07	(-2.59)	(-1.49)	(-1.83)	(-1.18)
NolexB-top	NolexB-bottom	**9.54****	**2.26**	11.09**	2.56	2.64	0.63	(-0.56)	(-0.16)
Lex4000-top	Lex4000-medium	**8.11****	**2.36**	2.63	0.84	5.17**	1.11	7.24**	2.75
Lex4000-top	Lex4000-bottom	**2.91**	**0.95**	4.39*	1.02	0.74	0.13	1.08	1.14
NolexA-top	NolexB-top	**-6.74****	**-2.47**	-7.31**	-2.47	4.97**	1.00	-4.51*	-2.86
NolexA-top	Lex4000-top	**-3.49**	**-1.37**	-1.28	-0.34	-0.55	-0.20	-4.39*	-4.40
NolexB-top	Lex4000-top	**1.21**	**0.45**	4.56*	1.49	-2.13	-0.07	-0.99	-0.56

Table 8.7: t-values of planned comparisons for Frequency

		TOTAL ($n = 20$)		**Nouns** ($n = 11$)		**Verbs** ($n = 4$)		**Adj/Adv** ($n = 6$)	
Category A	**Category B**	**By-Participt.**	**By-Targetwd.**	**By-Participt.**	**By-Targetwd.**	**By-Participt.**	**By-Targetwd.**	**By-Participt.**	**By-Targetwd.**
NolexA-top	NolexA-medium	**(-4.12*)**	**(-1.05)**	(-1.86)	(-0.36)	(-1.42)	(-0.53)	(-4.59**)	(-0.79)
NolexA-top	NolexA-bottom	**3.93***	**0.81**	(-0.05)	-0.14	3.46*	1.40	1.38	0.26
NolexB-top	NolexB-medium	**2.56**	**0.90**	(-1.48)	(-0.52)	3.47*	1.08	3.97**	1.16
NolexB-top	NolexB-bottom	**9.54****	**2.26**	1.25	0.40	7.73**	2.07	5.93**	1.30
Lex4000-top	Lex4000-medium	**8.11****	**2.36**	2.81*	0.96	4.54**	1.49	2.72	1.48
Lex4000-top	Lex4000-bottom	**2.91**	**0.95**	1.21	0.41	2.40	0.92	0.74	0.26
NolexA-top	NolexB-top	**-6.74****	**-2.47**	-2.08	-0.48	-3.89**	-2.23	-5.41**	-1.95
NolexA-top	Lex4000-top	**-3.49**	**-1.37**	-3.61**	-2.35	-1.36	-0.97	-0.41	-0.14
NolexB-top	Lex4000-top	**1.21**	**0.45**	-2.48	-1.42	0.79	0.51	4.35**	1.10

for nouns compared to all target words in that top-ranked examples are rated significantly higher than the bottom-ranked sentences but not the medium-ranked ones (for all target words, the picture is reversed). The only notable difference with respect to inter-model comparisons is that, compared to all target words, the Nolex-B model performs significantly better than the Lex4000 model for top-ranked sentences. As far as verbs and adverbs/adjectives are concerned, their considerably weaker performance is confirmed by the planned comparisons, both by the fewer number of significant t-values with the expected sign[10] and the increased number of t-values with the opposite sign.

An inspection of table 8.7 (for frequency categories) corroborates that high-frequency target words perform notably worse than medium- and low-frequency ones, both in terms of fewer number of significant t-values in the expected direction, and the increased number of t-values with the "wrong sign". Overall, both the medium- and low-frequency word results are fairly close to those of the total set of target words, both with respect to the intra-model comparisons and the significantly higher ratings of top-ranked sentences of the Nolex-B *vs* the Nolex-A model. For the Nolex-A and Lex4000 models, the medium-frequency words perform slightly better than the low-frequency ones with respect to the intra-model comparisons.

8.3 Evaluation Study II with Students

8.3.1 Participants

For the evaluation study with students, participants were sought among intermediate-to-advanced level students of L2 German. 34 students (both from German departments of British universities and the IALS language teaching institute) participated in the study.[11] 31 out of the 34 participants were native speakers of English.[12] Every participant was entered into a prize draw in return for their participation.

[10] However, it should be born in mind that due to the small n for both verbs and adverbs/adjectives, the significance levels are less indicative of the models' performances than they are for nouns.

[11] Data from another participant were discarded due to incomplete ratings.

[12] The native languages of the remaining participants were Hungarian, French, and unknown, respectively.

8.3.2 Materials

The materials consisted of a questionnaire that was largely identical to the questionnaire used for the evaluation study with teachers, but contained an additional section (included before the main body of the questionnaire) asking students to provide familiarity ratings for the set of 20 target words. The questionnaire was distributed both as a hardcopy and as an online, web-based questionnaire form. The same set of target words and example sentences as in the teacher evaluation study was used. As in that study, both the order of target words and the order of example sentences were randomized for each participant.

8.3.3 Procedure

The procedure used for the evaluation study with students was identical to the procedure used for the evaluation study with teachers, except for the following differences: participants were asked to (a) state their native language; (b) indicate their level of familiarity with each of the target words before proceeding to rate the example sentences for each target word; (c) indicate all unknown words in the original sentence contexts.[13]

For the familiarity ratings of the target words, a scale of 1 to 5 was used that largely[14] corresponds to the Vocabulary Knowledge Scale discussed in e.g. (Read, 2000) and uses the following five steps, or categories:

1 I don't remember having seen this word before, and I don't know what it means.
2 I have seen this word before but I don't know what it means.
3 I have seen this word before, and/or I think I know what it means.
4 I know this word.
5 I know this word and can use it in a sentence.

[13]The familiarity ratings were intended for a break down of the results into 'unknown' and 'known' target word categories. However, it was subsequently decided not to use these data, as a meaningful t-test analysis could not be carried out due to the considerable variability of familiarity ratings both across participants and target words.

[14]Minor differences in the wording of categories 1 and 3 are intended to reflect that the meaning of German compounds can often be guessed from their known constituent parts even if they have never been encountered before.

8.3.4 Results

8.3.4.1 General Results

As in section 8.2.4, a cursory inspection of the mean ratings across participants and target words (see table 8.8) yields first indications for the analysis (the corresponding teacher data from table 8.1 are given in brackets for comparison).

Table 8.8: Mean ratings across participants and target words

Model Type	Model Level	Mean	Std. Err.
Nolex-A	top	5.13 (3.93)	0.09 (0.15)
	medium	5.05 (4.52)	0.08 (0.15)
	bottom	4.74 (3.46)	0.08 (0.14)
Nolex-B	top	5.41 (4.70)	0.09 (0.16)
	medium	4.81 (4.28)	0.08 (0.15)
	bottom	4.74 (3.46)	0.08 (0.14)
Lex4000	top	5.38 (4.52)	0.09 (0.17)
	medium	4.84 (3.59)	0.09 (0.16)
	bottom	4.89 (4.07)	0.09 (0.16)
Dictionary	with penalty	4.92 (4.61)	0.11 (0.18)
Dictionary	no penalty	5.61 (5.24)	0.10 (0.18)
Teacher		6.66 (6.43)	0.08 (0.14)

Aside from the general observations that (a) student rating levels are significantly higher across all model categories compared to the teacher study;[15](b) the Nolex-B model (without the interaction term) is confirmed as the best-performing model overall, and (c) teacher examples still outperform all other categories, the following differences are particularly notable:

1. While dictionary examples with no penalty for missing entries are still rated consistently higher than the models' top-ranked examples, the dictionary examples with penalty are outperformed by the top-ranked examples of all models;

[15]The differences for all model categories are significant at the 1% level except for NolexA-medium and NolexB-medium which are significant at the 5% level; the differences for the dictionary and teacher examples are below significance level.

2. Both Nolex-A and Nolex-B models now exhibit internal consistency in terms of medium-ranked examples achieving lower mean ratings than top-ranked examples but higher mean ratings than bottom-ranked ones (in the teacher evaluation study, the Nolex-A model did not provide consistent calibration).

In order to analyze the student data in greater detail and test the hypotheses outlined at the beginning of the chapter, the same set of planned comparisons as the ones described in the previous section was carried out on the data. Again, the planned comparisons consisted of the paired t-test and relied on the Bonferroni-adjusted alphas. The results are presented in table 8.9 for both the by-participant and by-target word analysis. The table shows the respective t-values together with their associated significance levels: t-values significant at α_1 (two-tailed) are marked "*", and t-values significant at α_2 (two-tailed) are marked "**". T-values with the "wrong sign" (i.e. the difference between the respective mean ratings is in the opposite direction than expected) are given in brackets. The dictionary ratings with penalty are marked †, those with no penalty are marked ‡.

In keeping with the corresponding teacher data analysis, the paired t-test analysis reveals the teacher examples to be rated significantly higher than all of the top-ranked model selections, in both the by-participant and by-target word analysis.

In contrast to the teacher data analysis, where no significant difference could be found when comparing the dictionary example ratings involving the penalty for missing values to the ratings for the top-ranked model selections, the t-test analysis for the student data reveals the top-ranked examples of both the Nolex-B- and Lex4000 model to be rated significantly higher than the dictionary examples. For the non-penalized dictionary example ratings, the picture is very similar to the one shown by the teacher data analysis: the only significant difference is that between the dictionary examples (higher) and the Nolex-A top examples (lower) in the by-participant analysis.

With respect to intra-model comparisons, the results are even more encouraging than the ones found in the teacher data analysis. For the Nolex-A model, instead of the previously significant difference between the top- and medium-ranked sentences "in the wrong direction", top-ranked sentences are now ranked *higher* (albeit non-significantly so) than medium-ranked sentences, which in turn are rated significantly higher (at $p = 0.01$ rather than $p = 0.05$) than the bottom-ranked examples in the by-participant analysis. Again in the by-participant analysis, the Nolex-B model, which has maintained its consistent internal ordering, now exhibits a significant difference between the top- and medium-ranked examples. The Lex4000 model exhibits reduced

Table 8.9: t-values of planned comparisons

Category A	**Category B**	**By-Participant**	**By-Target Word**
Teacher	NolexA-top	-13.33**	-4.82**
Teacher	NolexB-top	-11.05**	-4.35**
Teacher	Lex4000-top	-12.89**	-4.24**
Dictionary†	NolexA-top	(1.95)	0.52
Dictionary†	NolexB-top	(4.74)**	(1.31)
Dictionary†	Lex4000-top	(5.17)**	(1.08)
Dictionary‡	NolexA-top	-2.35	-0.86
Dictionary‡	NolexB-top	(0.38)	(0.14)
Dictionary‡	Lex4000-top	-0.48	-0.14
NolexA-top	NolexA-medium	0.79	0.21
NolexA-top	NolexA-bottom	4.47**	1.16
NolexB-top	NolexB-medium	6.31**	2.33
NolexB-top	NolexB-bottom	7.28**	2.05
Lex4000-top	Lex4000-medium	6.39**	2.03
Lex4000-top	Lex4000-bottom	5.89**	1.46
NolexA-top	NolexB-top	-4.96**	-1.82
NolexA-top	Lex4000-top	-3.26*	-0.91
NolexB-top	Lex4000-top	0.47	0.11

inconsistency in its internal ordering and now features a significant difference between the top- and bottom-ranked examples (in the by-participant analysis).

The three inter-model comparisons show that, for the student data, the Nolex-A top-ranked examples are significantly outperformed not only by the Nolex-B top-ranked examples (as in the teacher data analysis), but by the Lex4000 top-ranked sentences as well (in the by-participant analysis).

As was the case for the teacher data analysis, in general, significance levels are reached much more readily in the by-participant analysis. This observation is not surprising, as — compared to the teacher evaluation — the number of target words remained the same ($n = 20$), while the number of participants increased from 14 to 34.

8.3.4.2 Teacher *vs* Student Mean Rating Correlations

An interesting question concerns the extent to which student ratings of the helpfulness of example sentences correspond to the respective teacher ratings for the same set of examples across the different categories.

In order to answer this question, a correlation analysis has been carried out on the teacher and student aggregated by-target word mean ratings. The results are presented in table 8.10 (all correlations are significant at $p < .01$).

Table 8.10: Teacher *vs* Student Correlations of Mean Ratings (By-Target word)

By-Target Word	**Corr.**	**By-Target Word**	**Corr.**
NolexA-top	0.87	Lex4000-top	0.94
NolexA-medium	0.76	Lex4000-medium	0.86
NolexA-bottom	0.59	Lex4000-bottom	0.86
NolexB-top	0.89	Dictionary (no penalty)	0.81
NolexB-medium	0.59	Teacher	0.55
NolexB-bottom	0.59		

Table 8.10 shows that teacher and student ratings correlate very highly for the models' top-ranked sentences (from $r = 0.87$ to $r = 0.94$) and also (to a slightly lesser degree) for dictionary examples ($r = 0.81$). Comparing the three models against each other, the Lex4000 model is the only model where student *vs* teacher correlations are very high across all categories (from $r = 0.86$ to $r = 0.94$), while the two Nolex mod-

els exhibit only moderate to moderately high correlations for the medium and bottom-ranked examples (from $r = 0.59$ to $r = 0.76$). Finally, correlations in the teacher example category are notable for being the lowest overall, while still in the moderate range ($r = 0.55$).

8.4 Discussion of the Evaluation Studies

The results of the empirical evaluation with teachers are encouraging with respect to the questions and hypotheses outlined in section 8.1.

Addressing the comparison of the models' preferred examples with the gold standard of teacher examples first, the analysis of both the teacher and the student ratings clearly confirms the expected result of teacher examples being consistently superior to even the best-ranked selections of the models. Almost all of the differences between teacher examples and top-ranked model selections were shown to be significant in both the by-participant and by-target word analysis. The only exception to this is the Teacher *vs* Nolex-B-top difference in the by-target word analysis of the teacher data, which was too small to reach significance level in the teacher data analysis. However, this result has not been confirmed by the student data anlysis; furthermore, since the by-target word analysis appears to be considerably less indicative of the models' performance than the by-participant analysis due to the lack of inter-target word consistency noted above, one should not necessarily take this as an indication that the top selections of the Nolex-B model perform on the same level as teacher examples.

Turning to the comparison of dictionary examples and the models' top selections, the results are very encouraging with respect to the teacher evaluation study, in that no significant difference between the ratings for dictionary examples and the models' top-ranked examples could be found. Even in the analysis slightly biased in favor of dictionary examples (the non-penalty version where only the available examples are considered), the only significant (at the 5% level) difference found was in the by-participant analysis in comparison to the Nolex-A model (which, as has been shown above, is outperformed by the Nolex-B model).

The results regarding dictionary examples are even more encouraging with regards to the student evaluation study. Here, not only could no significant differences be found between all models' top-ranked examples and the dictionary examples, but for dictionary examples involving a penalty for missing ratings, both the Nolex-B and the Lex4000 models' top-ranked examples are rated significantly *higher* than the dictio-

nary examples.

Questions (3) and (4) in section 8.1 (the inter-model and intra-model comparisons, respectively) together provide an insight into how the models perform relative to each other. The inter-model comparison can be broken down into two questions: (a) How do (in terms of their top-ranked sentences) the two Nolex models compare to each other?, and (b) How does (in terms of their top-ranked sentences) the better of the Nolex models compare to the model with the lexical complexity constraint? With respect to (a), the by-participant paired t-test analysis shows that the removal of the interaction term in the logistic regression equation (see chapter 7) does indeed improve the performance of the Nolex model in both the teacher and student evaluation. On the other hand, the lexical complexity constraint does not seem to affect the models' performance either way at least in the teacher data analysis, the slight caveat being that the non-significance found here (as well as for the dictionary *vs* model comparisons) should be seen in the context of the very conservative alpha (due to the high number of comparisons). However, in the student data analysis, the Lex4000 model's top-ranked examples are rated significantly higher (at the 5% level) than the Nolex-A model's top-ranked sentences, while the difference between NolexB-top and Lex4000-top remains insignificant.

Overall, the Nolex-B model (with no interaction component or lexical complexity constraint) outperforms both the Nolex-A and the Lex4000 model, since — in the teacher evaluation — it is the only model with both consistent internal ordering and significant differences between the model's top- and bottom-ranked sentences. In the student data analysis, this is true for both Nolex models; even more importantly, for both the Nolex-B and Lex4000 models, significant differences here exist not only between the top- and bottom ranked sentences, but also between the top- and *medium*-ranked sentences (in the by-participant analysis). Since, in addition, the Nolex-B model features consistent internal ranking of the categories, this means that — for the student evaluation study and in the by-participant analysis — at least one of the models has been shown to improve on a random selection of example sentences from corpora.

Although quite positive especially with respect to the student evaluation study, the results of this evaluation study should be accepted with caution. This is not so much because of the relatively small number of participants of the teacher evaluation study (only 14 teachers took part in the study), but rather because inter-item consistency appears to be quite low in spite of twenty target words being tested. This means that the quite encouraging results of the by-participant analysis could not be confirmed by

the by-target word analysis.

While the analysis of the subgroups according to parts-of-speech and frequency groupings (which was only carried out for the teacher data analysis) provides first indications that nouns perform significantly better than verbs, adjectives and adverbs, and high-frequency words do not perform on the same level as medium or low-frequency words, these results need to be validated by a larger-scale study that remedies the following shortcomings of the present study: (a) a much larger number of target words needs to be tested; (b) the target words should be selected in a balanced way that takes into account not only the factors tested above (parts-of-speech and frequency), but also other factors potentially influencing the rating of example sentences. A look at table 8.2 showing the mean ratings for the 20 target words suggests that idiosyncrasies of the target word selection may have influenced the results, and that one such additional factor is what might loosely be described as "How amenable is the target word to an explanation by an example sentence, as compared to a definition?". For instance, the lowest-rated target word in table 8.2, *Currywurst* (curry sausage), being a very specific food item, seems particularly suited to a definitional approach as opposed to an example sentence.[16] By way of contrast, the highest-rated target word, *Wassertemperatur* (water temperature) seems reasonably well explained by most randomly-selected sentences, since most sentences containing *Wassertemperatur* are likely to feature both references to some form of water and temperature-indicating words like *degree* or *Celsius*.

Several other design limitations of the evaluation study that should be addressed by any follow-up studies can be summarized as follows:

- Frequency in German is not necessarily the best indicator of word difficulty for learners of L2 German because of the prolific compounding and may need to be replaced may a more suitable but difficult-to-define construct; for example, compounds such as *Wassertemperatur* (water temperature) are relatively infrequent yet easily guessable (at least for learners with L1 English).

- For polysemous words, even though teachers were told that all target words were used in the examples *in roughly the same sense* as in the original sentence, it is possible that fine-grained sense distinctions have influenced the participants' ratings; this relates back to the general problem of word sense disambiguation

[16] As several teachers that participated in the study have remarked after completion of the questionnaire.

addressed in chapter 6.

- Even though participants were told that they should not let their ratings be influenced by the lack of context in some cases, it is possible that they did not always remember this instruction for all ratings. Any follow-up validation study should address this problem by controlling for lack of context (as measured by e.g. anaphora in the sentence). Furthermore, a topic of an example sentence may appear diffcult to grasp not only because of lack of context but also because knowledge of the subject matter is lacking, and it may prove difficult to draw a clear dividing line between the two.

- No clear instructions were given to teachers as to the exact level that they should assume their hypothetical students to be at, other than that they should assume a sufficiently advanced knowledge of syntax. This was both for practical and design reasons (level names and descriptions vary across teaching institutions; as few as possible restrictions should be placed on the number of potentially participating teachers and the number of target words tested). However, any follow-up study should attempt to control the assumed vocabulary level more tightly, both for teachers and students as participants.

The above caveats and limitations notwithstanding, the results of the current evaluation nevertheless suggest that at least one of the models tested — Nolex-B with no interaction term and no lexical complexity constraint — reflects a substantial number of the considerations that teachers employ in their selection and judgment of example sentences, and that its top-ranked example sentences — in the teachers' judgment — are similarly helpful to students as examples provided by latter-day dictionaries of L2 German.

Even more significantly, the evaluation of the models with students shows that the Nolex-B model selects examples that students of L2 German perceive as significantly more helpful than randomly selected corpus examples, and that the top-ranked examples of two of the models tested — Nolex-B and Lex4000 — significantly outperform dictionary examples in terms of perceived helpfulness to students when a penalty for missing entries is included.

This is an encouraging finding considering the very limited number of dictionary examples compared to potential corpus examples. The results also indicate that the inclusion of a lexical complexity constraint, i.e. the exclusion of examples with too

difficult vocabulary, does not significantly enhance or diminish the helpfulness of the example sentences.

8.5 Summary

In this chapter, three models presented in the current thesis were evaluated by both teachers and students of L2 German. The empirical evaluation involved example sentences representing 11 different categories for 20 target words being rated by both experienced teachers and intermediate-to-advanced level students of L2 German. Even though the results need to be considered preliminary due to the limited number of target words and limitations of the study discussed above, they are encouraging. For the teacher evaluation study, the findings indicate that one of the models tested performs not only on the same level as dictionary examples for the top-ranked selections, but also provides a ranking of potential examples that is roughly in line with that of experienced teachers of L2 German. The student evaluation has confirmed these results and improves on the findings of the teacher evaluation in at least two important respects: the best-performing model of the teacher evaluation selects examples that, in the students' evaluation, significantly outperform both randomly selected corpus examples and dictionary examples (when a penalty for missing entries is included).

Chapter 9

Discussion and Conclusions

9.1 Summary of the Research

The current thesis examined criteria for the selection of example sentences for difficult or unknown words (target words) in reading texts for students of German as a Second Language. The intended use of the examples is within the context of an Intelligent Computer-Aided Language Learning (ICALL) Vocabulary Learning System, where students can choose among several explanation options for target words. Aside from example sentences, these are envisaged to be pictorial or textual glosses (such as translations, definitions, or paraphrases).

There has been extensive research in Second Language Learning and ICALL quarters on the effectiveness of different types of glosses for incidental L2 vocabulary learning (see Yoshii (2006)) for an overview). By way of contrast, example sentences have been virtually unexplored as an explanation option in ICALL Vocabulary Learning Environments.

The selection of example sentences (ES) for the to-be-developed Vocabulary Learning Environment differs from the selection of textual glosses such as definitions and translations in that the former is a non-trivial task, given the vast range of potential ES (the only constraint for possible candidates of ES of a target word is that the target word be contained in the examples).

It was argued in chapter 1 that the approach of restricting the source of ES to dictionaries has several disadvantages: first and foremost, the number of available dictionary examples for a given target word is severely limited. However, a learner clearly benefits from being exposed to a wide range of example usages of the target word. Second, while it may seem reasonable to presume that a dictionary example is more helpful

to the language learner than a randomly-picked corpus example, the selection criteria for the dictionary ES are often opaque. We saw in chapter 2 that much of the lexicographic discussion on example sentences focussed on the question of whether to use authentic (corpus-*attested*) examples, or examples that are only corpus-*oriented* (i.e. possibly invented by the lexicographer as he sees fit, using corpus occurrences only as a guideline). Learner dictionaries in particular tend to differ to a considerable extent in the leeway they allow the lexicographer in this respect. Third, ES taken from dictionaries do not take into account the reading context in which the word was encountered. However, a language learner reading an L2 text may well benefit from an example that is semantically similar to the original sentence containing the target word.

The survey of the lexicographic discussion on ES in chapter 2 suggested that the question of most interest for this study — *What makes a good, i.e. helpful, ES for an L2 learner?* — has either been largely neglected, or reduced to the above-mentioned discussion of authentic *vs* invented examples. Given this situation, the approach taken in this thesis was a pedagogic, teacher-centered one, namely a first exploratory investigation of the question *What makes a good, i.e. helpful, ES for an L2 learner from an L2 teacher's perspective?*

Central to this thesis was the hypothesis that a random selection of ES from a suitable corpus could be improved by a guided selection process that takes into account characteristics of examples perceived to be *helpful* by experienced teachers of L2 German. In order to investigate this issue, an empirical study was conducted, the purpose of which was twofold: first, to elicit ES in the form of invented examples from experienced teachers of L2 German for unknown or difficult target words in a reading text that they believed to be most helpful for illustrating or clarifying the meaning of the word; and second, to gather explanations from the teachers as to the reasons why they considered their examples particularly helpful. The teachers' explanations were analysed in chapter 3 and provided a basis for the analysis of the teachers' ES along several dimesions suggested by the explanations: in particular, these were criteria of context (analyzed via semantic sentence similarity in chapter 5), the reduction of syntactic complexity (chapter 4), and specific lexical choices relating to both words in the ES, and the morphological form of the target word itself (chapter 6).

In order to analyze whether the syntactic complexity of the teacher-provided ES had been significantly reduced compared to the corresponding original sentences (OS), chapter 4 first considered several measures of syntactic complexity suggested in the literature; these were then empirically evaluated on the basis of a correlation analysis

with native speaker judgments of syntactic complexity. Of the measures tested, the simplest possible measure — sentence length — was found to yield the best correlations with native speaker judgments, and was therefore used to measure the reduction of syntactic complexity from OS to ES. Using a paired t-test, this analysis revealed that syntactic complexity was significantly reduced in the teachers' examples compared to the OS. The chapter also motivated the use of syntactic complexity as a pre-filter for the overall model of the teacher criteria later developed in chapter 7, using the best-fit linear function as a predictor for the syntactic complexity for candidate examples.

In chapter 5, the context criterion was looked into via a comparison of the semantic sentence similarity of the OS-ES pairs in the teacher data. The chapter first surveyed several measures of sentence similarity suggested in the literature, with lexical overlap and Latent Semantic Analysis (LSA) being selected as analysis measures to be evaluated for the study at hand. Of these measures, LSA was found to yield the best correlations to human sentence similarity judgments, which had been elicited in a web-based empirical study. This motivated the use of LSA as the measure of sentence similarity for the teacher data. More specifically, LSA was used to address the question of whether the degree of sentence similarity found in the teacher data was significantly higher than that found in randomly selected sentence pairs. This was indeed found to be the case, motivating the inclusion of sentence similarity as a predictor variable to be included in the model of teacher criteria for the selection of examples (chapter 7). A significant "by-product" of the analysis presented in chapter 5 was the finding that standard LSA, used for German with a suitably big training corpus (at least ca. 60 MB), can achieve remarkably high correlations with human sentence similarity judgments. This result clearly exceeds previously reported correlation levels for standard, non-lemmatized LSA without syntactic or relational components.

Turning to specific lexical choices teachers may have made in their ES, chapter 6 was concerned with the investigation of three types of lexical choices: paradigmatic lexical relations, significant co-occurrences, and the morphological form of the target word itself. The main finding in the chapter was that both paradigmatic relations and significant co-occurrences were significant factors in the teacher examples, while morphological choices relating to the target word form were not. As a side issue in relation to word similarity measures, LSA (representing vector-space measures) and LC-IR (representing statistical web-based approaches) were compared on the tasks of a multiple-choice lexical relation test, and a correlation analysis with native speaker similarity ratings of German noun pairs. The result was that their performance in both

tests was inconclusive and less-than-optimal for the task at hand. It was therefore concluded that further research was needed before the automatization of the task of measuring word similarity could be considered for any future implementation of the model of the teacher criteria developed in chapter 7.

In chapter 7, models based on the above-mentioned factors that had proved significant in the selection of the teachers' examples were developed using logistic regression anaylsis. The models were evaluated both with and without the inclusion of a lexical complexity constraint (which took the form of excluding content words outside the enhanced Basic German Vocabulary of ca. 4,000 words); two versions of the model without the lexical complexity constraint (Nolex-A and Nolex-B) were tested, motivated by the spurious behavior of the original Nolex-A model for certain value ranges. The logistic regression analysis of the three models served as the basis for the evaluation analysis of the models described in chapter 8.

For the evaluation of the logistic regression models, their output was submitted to both experienced teachers and intermediate-to-advanced level students of L2 German for judgment of their respective helpfulness. The evaluation addressed the question of how the models' preferred examples compared with (a) the gold standard of ES provided by an experienced teacher of L2 German; (b) examples provided by suitable dictionaries; (c) each other. The question of whether the models provided consistent internal ordering, and whether the the models' top-ranked examples were rated as significantly more helpful than both random corpus selections and bottom-ranked examples, was also considered in the evaluation. The results, which need to be considered preliminary due to the limited number of target words and design limitations, were found to be encouraging: for the teacher evaluation study, they indicate that one of the models tested — Nolex-B — performs not only on the same level as dictionary examples for the top-ranked selections, but also provides a ranking of potential examples that is roughly in line with that of experienced teachers of L2 German. The student evaluation has confirmed these results and improves on the findings of the teacher evaluation in at least two important respects: the best-performing model (Nolex-B) of the teacher evaluation selects examples that, in the students' evaluation, significantly outperform both randomly selected corpus examples and dictionary examples (when a penalty for missing entries is included).

9.2 Critical Remarks

The issues researched in this dissertation are complex on various levels, and it is essential to emphasize that various simplifying assumptions were made in order to narrow the scope of the investigation down to a feasible level. Many methodological decisions that resulted from this are no doubt open to a considerable amount of criticism. By the same token, it is important to stress that due to the various limitations and potential criticisms that could be leveled against various aspects of the work presented in this thesis, the resulting model has to be considered a preliminary first approximation of a more refined model of teacher selection criteria that remedies the weaknesses of the approach discussed below. At any rate, it is clear that the model presented requires substantial further development, before it can be considered as the basis for the example selection component of any "real-world" implementation of the envisaged Vocabulary Learning Environment. In the remainder of this section, some of the critical comments regarding the present work are addressed.

Should the approach of the thesis in general have been learner-centered rather than teacher-centered? The work presented in this dissertation was motivated in the main by the question "What makes a good, i.e. helpful, ES for an L2 learner *from an L2 teacher's perspective*?" This pedagogic, teacher-centered perspective on the issue informed both the exploratory empirical study that provided the data to be analyzed (in the form of teacher-invented examples and their corresponding explanations), and the first evaluation study of the model (which also used teachers as participants). "Teacher-modeling" appeared preferable to "student modeling" in the context of a first exploration of the issue of what criteria should be applied to the selection of example sentences for the following reasons: (a) as regards the exploratory study, experienced teachers are — in contrast to language learners — able to create or "invent" helpful (according to their judgment) example sentences; arguably, they are also able to a much greater extent than language learners to verbalize what makes an example helpful; (b) as far as the evaluation is concerned, focussing only on a student evaluation would not have permitted the assessment of the models' success in modeling the teachers' criteria. The evaluation study shows that the teachers' judgments of the helpfulness of example sentences tend to correlate very highly with student ratings. Furthermore, the student evaluation study demonstrates the helpfulness of the models' top selections to students, even though further studies should be carried out with students to assess the

actual learning effect (see section 9.3).

Should the teachers have been asked in the exploratory study to select among corpus examples, rather than create their own? It might be argued that having teachers select examples would have been preferable to asking them to invent examples for the following reasons: (a) since the developed models select among 'authentic' corpus examples, the yardstick against which they are measured ought to be commensurate, i.e. preferred 'authentic' examples should have been used rather than invented examples; (b) teachers vary in their ability to be creative so that the implied presumption of 'ideal' teacher examples as the gold standard is questionable; (c) not adopting the corpus-selection approach in the exploratory teacher study meant that no negative examples were available as input for the logistic regression model; these examples had to be 'artificially' created on the basis of the analysis of the positive examples instead.

Addressing point (a) first, the stance taken in this thesis is that the distinction between 'authentic' and 'invented' examples is largely irrelevant for language learning purposes (see discussion in chapter 2). Given the practical unfeasability to rate more than a tiny fraction of existing authentic corpus examples, clearly only invented examples provide the opportunity to analyze example data that teachers consider maximally helpful. As for (b), it has to be conceded that teachers vary in their creativity with respect to providing examples; however, we believe that this caveat is a minor one, outweighed by the general advantage of the "invented example" approach mentioned above. What is more, the task of controlling for verbal creativity as a criterion for participant selection would have been difficult to achieve, and the corresponding overhead appears out-of-proportion vis-à-vis the potential benefit. Regarding (c), the detrimental impact of the chosen approach for the development of the logistic regression models has to be conceded (see also the point below); however, we believe it had to be accepted given the above-mentioned fundamental drawback of an "authentic example" approach for the exploratory study.

Was the approach used for deriving the negative input for the logistic regression model valid? As was mentioned above, the fact that only positive examples were available as input for the logistic regression model (in the form of invented teacher examples) meant that the required negative input had to be created 'artificially', as described in chapter 7. This is clearly a methodological weakness of the adopted regression modeling approach: it resulted in regression models that were skewed to an

extent, in the sense that the *range* of possible values of the predictor variables arguably gained undue precedence over the actual *relative importance* of the factors. As was noted in chapter 7, the most obvious result of this bias is the unduly large effect that the factor with the broadest range of values — sentence similarity, which was measured by vector cosines — exerted over the model ratings. This caused even small increases in sentence similarity to yield substantially improved ratings, even if the absolute similarity value remained well below the "high similarity" threshold established in chapter 5. Since this undesired behavior of the model is intrinsic to the "invented example" approach, we believe it had to be accepted for the current exploratory study for the reasons given above, but should be addressed in future improvements on the model (see point (2) in the following section).

Should we have concentrated on teacher participants who had target students from the exact same level of proficiency, rather than the comparatively vague concept of "advanced-level students"? Without doubt it would have been desirable to treat the target students' level of proficiency (especially with respect to reading proficiency and vocabulary knowledge) as a strictly controlled variable — rather than the less well-defined notion of "advanced-level" students for whom grammar does not present a major obstacle for reading comprehension — and use only teachers for the exploratory study and evaluation who had students from the exact same level of proficiency.

This would have been desirable mainly with respect to the treatment of lexical complexity as a potential factor in the analysis, rather than using the somewhat arbitrary — and quite conservative — lexical complexity constraint of the enhanced Basic German Vocabulary as a distinguishing factor for the two types of models developed in chapter 7. A more strictly controlled level of proficiency might also have had an impact on the the analysis of the reduction of syntactic complexity, and on the morphological choice of the target word form. On the other hand, it is difficult to imagine the factors of semantic similarity and lexical choices being affected by this aspect. However, two factors mitigated against a stricter control of the proficiency levels: first, insofar as a common reference scheme of proficiency levels exists (the Common Reference Levels of General Language and Reading Proficiency), it is arguably too coarse-grained for the purposes of this study (and even if it were not, the relevant questionnaire sections filled in by the teachers indicate that they find it difficult to assign their classes to just one category). Second, experienced teachers of L2 German willing to participate in

the study were in relatively scarce supply, so it would have been undesirable to reduce their number further by imposing additional selection restrictions.

Was the manual approach used for Word Sense Disambiguation (WSD) valid? An implementation of a WSD module was outside the scope of this thesis. Furthermore it is unclear whether automatic WSD would have been usable, since systems embodying such an approach have limited coverage and typically only deal with nouns. Given this situation, a manual approach guided by dictionary entries was taken; while a purely dictionary-driven approach would have been preferable due to its avoidance of a certain degree of arbitrariness that was inevitably present in the approach adopted, it was rejected due to the reasons given in section 6.2.

A related problematic aspect of the WSD treatment concerns the analysis of polysemous target words; due to the limited amount of polysemous target words contained in the teacher data, the OS to ES word sense transitions received only a cursory inspection in this thesis. This inspection motivated a 'liberal' manual approach to word sense selection that only ruled out clearly distinct word senses; this approach might well be proven untenable by a thorough analysis on the basis of target word data controlled for polysemy. However, as was pointed out in section 6.2, this is unlikely to be a serious practical problem for advanced learners as the issue of polysemy tends to decrease in importance for more infrequent (difficult) words. By a similar token, the related analysis of figurative *vs* literal usages in the OS/ES pairs should be based on more data to confirm the preliminary transition patterns indicated by the current teacher data.

Was the "one-rater" approach used for Selection of the Test Data Items for Syntactic Complexity and Sentence Similarity valid? Ideally, the selection for the test items should have been based on more than one rater to rule out possible rater bias in the case of sentence similarity test pairs; for the syntactic complexity test items, ideally the selection of test sentences should have controlled for syntactic complexity aspects that the tested measures are based on. However, the concomitant increased complexity of the selection design appeared to be disproportional to the gain in soundness for the purpose of analysing the teacher data. The cost of this decision is that no claims can be made about whether or not sentence length is the best measure of syntactic complexity for German in general.

It also ought to be pointed out that the current evaluation of syntactic complexity measures will not necessarily hold for any future extensions of the model that take

multi-word lexical items (such as verbs with separable prefixes) into account, since discontinuous syntactic constituents arguably increase the syntactic complexity of short sentences to a non-negligible extent.

Was the approach used in discarding teacher examples and explanations valid? Since the exclusion of teacher examples and explanations was based on introspection of the author, a slight degree of arbitrariness was inevitably present in these decisions as well. This is especially true for decisions as to when an example was considered too close to a definition, considering that no clear-cut dividing line exists between the two. In retrospect, these decisions would have been sounder had they been made by a panel of raters. However, it needs to be said that only a few examples could be considered problematic 'borderline' cases, so that the adopted approach is unlikely to have made a significant detrimental impact.

What are the implications of the dictionary-based, 'manual' approach for automatic example sentence selection? As has been noted in chapter 6, the choice of the dictionary-based, 'manual' approach to word similarity detection is not optimal as it ensures maximum precision, but offers only low recall (due to the insufficient coverage of existing lexical resources, and the fact a non-negligible amount of semantic relations are non-classical in nature). While its automatization is conceivable to the extent that the corresponding lexicographic resources are available in electronic form, it is evident that further research will be required to arrive at a measure of word similarity more satisfactory for the task at hand (see also the following section).

9.3 Further Work

The research presented in this thesis is intended as a first stepping stone towards modeling the criteria that are helpful to learners of L2 German in the context of reading texts in an ICALL Vocabulary Learning System. As such, it is inevitably lacking in certain respects and provides ample opportunities for further improvements and extensions, which relate to various levels of the work presented. In the remainder of this section, we provide a (non-exhaustive) list of suggestions for further work that could be conducted on the basis of this research.

1. **Performing further evaluation studies.** The evaluation of the model discussed in chapter 8 is quite encouraging and suggests that the approach taken in this thesis is valid at least as a first exploratory investigation into the selection of example sentences based on teachers' selection criteria. However, owing primarily to the small number of target words tested, inter-item consistency was apparently too low for the by-target word analysis to confirm the largely encouraging results of the by-participant analysis. Further evaluation studies could attempt to rectify this problem by both using a much larger test set of target words, and by taking factors into account that might influence the ratings of ES. As was pointed out in chapter 8, an important such factor may relate to point (4) below, i.e. the "suitability" of the target word for an example as opposed to e.g. a definition. Further evaluation studies could also address other design limitations of the current evaluation study that were summarized in section 8.4.

2. **Improving the logistic regression model with "natural" negative input.** As was discussed in the preceding section, the 'artificial' negative input selected for the regression analysis had unwelcome repercussion on the model behavior. Now that the current study has identified basic significant factors of the teacher criteria for ES selection, a follow-up study could use these as selection criteria for corpus examples to be presented to teachers, i.e. take the "rate authentic examples"-route for gathering new input data for the regression models. This would have the added benefit of eliminating the current restriction whereby the helpfulness of the examples is being evaluated only in respect to their *decoding* function (i.e. the extent to which they succeed in illustrating the meaning of the target words). In the lexicographic literature, 'authentic', corpus-attested examples are widely considered essential to meet the *encoding* needs of the language learner (Fox, 1987).

3. **Evaluating the actual learning effect of the models for students.** It has been cautioned above that, ultimately, the evaluation standard for future developments of the model should be how well they help students learn the unknown words. There are several dimensions of "success" that could be tested here (e.g. text comprehension, short-term word comprehension and long-term word retention). As has been mentioned in chapter 2, previous research (Mondria and Wit-De Boer, 1991) has shown that factors that are conducive to guessability and text comprehension may not be conducive to long-term retention.

4. **Testing for the respective usefulness of example sentences *vs* definitions and other explanation options for different types of target words.** The evaluation of the models in chapter 8 — as well as some of the teacher comments and explanations in the exploratory study — clearly suggested that some words are much less amenable to the example sentence treatment and would be better served by definitions (e.g. *Currywurst* (curry sausage)). This issue also lends itself to an empirical analysis, where the actual usage of the respective explanation options as well as their helpfulness for the students are investigated. This question may be related to another aspect of examples that was not considered in this thesis, namely the extent to which examples constrain the meaning of the target word (see the discussion of "forcing examples" in (7)). An interesting, wider-ranging question would be to compare the helpfulness of the different explanation options for both comprehension and retention (see Mondria (2003) for an investigation into the effects of "meaning-inferred" (e.g. examples) *vs* "meaning-given" (e.g. definitions, translations) methods on the retention of L2 word meanings).

5. **Considering the lexical complexity constraint as a model factor**. As the discussion in the preceding section has suggested, follow-up studies on teacher criteria that control the students' (vocabulary) level of proficiency more tightly would allow the investigation of lexical complexity as a factor to be included in the model, rather as the (for advanced students) arguably too conservative 'post-hoc' lexical complexity constraint used in this thesis.

6. **Extending the model to cover multi-word target words and idioms.** An obvious extension of the current model concerns the inclusion of multi-word lexical items, such as verbs with separable prefixes and idioms. The inclusion of these types of target words would necessitate the re-analysis of the reduction of syntactic complexity, as well as the analyses of morphological forms of the target words. It appears unlikely that the extension will have a significant effect on the remainder of the analysis dimensions.

7. **Extending the analysis to cover "forcing examples" as a selection criterion.** As was discussed in chapter 2, "forcing examples" are examples that constrain possible meanings of the target word to such an extent that only one plausible interpretation as to the target word's meaning remains. This aspect may be related to the "definitions *vs* example sentences" issue discussed above: it might be tentatively speculated that the scarcity of authentic examples that can be considered

forcing examples (or the difficulty of concocting an invented forcing example) can be taken as an indicator that definitions are better suited to the explanation of the respective target words than example sentences are.

8. **Automating Word Sense Disambiguation.** The implementation of a state-of-the-art WSD module is a pre-requisite to any future implementation of the models developed in this thesis (however see the caveat regarding WSD in the preceding section).

9. **Investigating teacher criteria with respect to different word senses.** It was pointed out in the preceding section that for polysemous words, the teacher criteria with respect to possible word sense transitions from OS to ES need to receive a more thorough analysis on the basis of target word data controlled for polysemy. By a similar token, in order to confirm the preliminary transition patterns indicated by the current teacher data, the analysis of figurative *vs* literal usage transitions in the OS/ES pairs should be based on more data controlled for frequency of figurative target word usage.

10. **Automating the word similarity measure.** As was discussed in the preceding section and chapter 6, the current dictionary-based approach to detecting word similarity is suboptimal and requires further research. Possible approaches to this problem include: (a) given that the analysis results for the larger test set (the multiple-choice lexical selection test) are quite encouraging for LC-IR, this measure could be tested on a larger and more comprehensive data set than the 57 noun pairs available for this study; (b) LSA could be extended with lemmatization, syntactic or relation information along the lines suggested in chapters 5 and 6; (c) once extended versions of GermaNet which have a coverage comparable to WordNet become available, lexical network-based approaches to word similarity not considered in this thesis could be analyzed.

11. **Improving the sentence similarity measure.** Even though standard LSA achieved correlations with human sentence similarity judgments that are remarkably high, it is still possible that enhanced versions of LSA (lemmatization, inclusion syntactic or relation information) could achieve an even better performance. This hypothesis is at least partially supported by Zipitria et al.'s (2006) findings that while lemmatization does not show notable improvements for a Romance, non-agglutinative language such as Spanish, it does yield significant improve-

ments for Basque (an agglutinative language); German arguably lies somewhere in between Spanish and Basque on the non-agglutinative to agglutinative scale. The potential benefit of incorporating relational or syntactic information into LSA is also suggested by Wiemer-Hastings's (2004) research in this area which shows that verbs play a predominant role in human sentence similarity judgments, while the role of syntactic subjects is largely ignored in such ratings.

9.4 Conclusions

The purpose of the research presented in this thesis was to investigate possible criteria that teachers employ in their selection of example sentences for students of L2 German reading a German language text with unknown or difficult target words. The focus of this research was specifically teacher-oriented in that teachers provided not only the example sentence data that formed the basis for the analysis presented in chapters 3 to 6, but also evaluated the models based on this analysis. Whether the models can be straightforwardly extended to cover multiple-word lexical items, whether their performance can be improved by the various extensions and improvements suggested, and whether they actually prove helpful to the intended target group – advanced-level learners of L2 German — remains to be seen. All in all, despite the various obvious limitations and simplified assumptions on which the models are based, the teacher evaluation has shown that one of the models fares well both as compared with dictionary examples, and in terms of providing a ranking of potential example sentences that roughly corresponds to that of experienced teachers of L2 German. Furthermore, the student evaluation has not only confirmed these results, but improved on them by significantly outperforming both random corpus selections and dictionary examples (with penalty for missing entries).

Appendix A

Sample Questionnaire for Teacher Study

TEACHER QUESTIONNAIRE: CHOOSING EXAMPLE SENTENCES FOR DIFFICULT TARGET WORDS IN A GERMAN TEXT

Thanks for participating in this study about the choice of example sentences for difficult target words in German newspaper/magazine text. During the following hour, you will be asked to (a) identify and classify difficult target words in the text provided; (b) give an example sentence for each target word; and (c) explain your criteria for choosing your example sentences.

Before you set about this task, please provide the following background information:

TEACHER NAME: _____

PART I: TEACHER-RELATED BACKGROUND INFORMATION

1 How long have you been teaching L2 German? _____

2 If you are teaching any other L2 languages, then please indicate which languages, and for how long you have been teaching them:

LANGUAGE	NUMBER OF YEARS

3 What is your native language? _____

4 When a student comes across an unknown word, how often would you illustrate the word with an example sentence?

◯ never ◯ rarely ◯ sometimes ◯ often ◯ always

PART II: CLASS-RELATED BACKGROUND INFORMATION

For each of the classes you are teaching, please provide the following information:

CLASS TITLE	NUMBER OF STUDENTS	BREAKDOWN BY L1	GENERAL LANGUAGE PROFICIENCY*	READING PROFICIENCY*
Example Class Name	**17**	**English: 12** **Spanish: 5**	**B2**	**B2-C1**

∗ For general language and reading proficiency, please rate your classes in terms of the Common Reference Levels (Council of Europe), an adapted version of which is provided below:

	General Language Proficiency	Reading Proficiency
A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of need of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.	Can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.	Can read very short, simple texts. Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and can understand short simple personal letters.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.	Can understand texts that consist mainly of high frequency everyday or job-related language. Can understand the description of events, feelings and wishes in personal letters.
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	Can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes and viewpoints. Can understand contemporary literary prose.
C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.	Can understand long and complex factual and literary texts, appreciating distinctions of style. Can understand specialised articles and longer technical instructions, even when they do not relate to his/her field.
C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in complex situations.	Can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

PART III: TARGET WORDS AND EXAMPLE SENTENCES

General Overview

Imagine teaching one of your L2 German classes (preferably the one nearest B2 level). Please indicate here which class you have selected: _____.

Please carefully read the following instructions, and the reading text provided.

1) Please identify all words (or lexical units [e.g. special collocations, idioms] such as *ins Gras beißen*) that might be unknown (to some degree) to the average learner in the selected group. For each word, please provide a numbered index by writing 1,2,3 etc. above it in the reading text.

2) After having completed 1), please supply the following information in the matching section of the form labelled PART III - FORM:

A Insert the word;

B Classify the word using the following rating scale:

(a) unfamiliar but could guess from world/topic knowledge

(b) unfamiliar but could guess from surrounding context

(c) completely unfamiliar, or unfamiliar in this particular meaning

(d) unfamiliar in form
(e.g. irregular past tense, separated verb prefix at end of sentence)

(e) unfamiliar compound: know all parts but not compound meaning

(f) unfamiliar compound: know some but not all parts

(g) unfamiliar for other reason: please specify

C For each identified word, please give the best example sentence that you can think of to illustrate the meaning of the word, taking into account:

(a) the vocabulary and grammatical knowledge level of your students;

(b) the surrounding context of the target word in the reading text;

(c) general interest areas and world knowledge of your students.

NB: Please provide **exactly one** example sentence for each target word (even if you think that another explanation option - e.g. definition, paraphrase - would be more helpful for the given word). **Please make sure that the target word is used in your sentence.**

D For each example sentence, explain why you chose this sentence as a helpful sentence for illustrating the meaning of the word. What criteria did you use? Is this example sentence specific to this text? If so, in what ways? If not, why not? Please be as detailed and specific in your answer as possible.

PART III - FORM

INDEX #	TARGET WORD	CLASSIFICATION (a - g)
EXAMPLE	**Trinkgeld**	**e**

Example sentence: *In diesem Restaurant ist der Service im Preis inbegriffen, aber es ist trotzdem üblich, dem Kellner ein* **Trinkgeld** *zu geben.*

Explanation/Criteria: *[your explanation/criteria here]*

****

****

#	TARGET WORD	CLASSIFICATION (a - g)
1		

Example sentence: _____

_____

Explanation/Criteria: _____

_____

_____

_____

_____

_____

#	TARGET WORD	CLASSIFICATION (a - g)
2		

Example sentence: _____

_____

Explanation/Criteria: _____

_____

_____

_____

_____

_____

#	TARGET WORD	CLASSIFICATION (a - g)
3		

Example sentence: _____

_____

Explanation/Criteria: _____

_____

_____

_____

_____

_____

Appendix B

Teacher Examples (ES) with their corresponding Original Sentences (OS)

In the following, all 243 teacher examples (ES) retained for the analysis are listed together with their corresponding Original Sentences (OS) in the reading texts.

Target Word #1: findig (resourceful)

OS: Vom findigen Mönch, dem die Menschheit wohl die Brezel verdankt, ist leider weder der Name noch der genaue Ort seiner Erfindung bekannt.

ES: Als Erfinder war er findig, aber als Mensch unsympathisch.

Target Word # 2: Ansporn (incentive)

OS: Fleißige Klosterschüler bekamen sie als Ansporn für das Lernen neuer Gebete.

ES: Kindern gibt man kleine Belohnungen als Ansporn zum Lernen.

Target Word # 3: strafverschärfend (leading to a more severe punishment)

OS: Strafverschärfend fiel ins Gewicht, dass sie für den Teig gutes Mehl verwandt hatten, während sie das Brot aus pampiger Kleie an die Weissen verscherbelten.

ES: Strafverschärfend fiel ins Gewicht, dass der Verbrecher keine Reue gezeigt hatte.

Target Word # 4: verscherbeln (to fob off)

OS: Strafverschärfend fiel ins Gewicht, dass sie für den Teig gutes Mehl verwandt hatten, während sie das Brot aus pampiger Kleie an die Weißen verscherbelten.

ES: Der fliegende Händler verscherbelte schlechte Waren an seine Kunden.

Target Word # 5: Verzehr (consumption)

OS: Statistisch gesehen besteht allerdings bis Anfang Februar für Amerikaner erhöhte Lebensgefahr beim Verzehr, dann nämlich endet die Football-Saison.

ES: Beim Verzehr von Brezeln erhöht sich die Gefahr eines Erstickungsanfalls.

Target Word # 6: nach Belieben (at will)

OS: Die Baltimore Ravens, die Titelverteidiger mit ihrer bärenstarken Abwehr, beherrschten die Miami Dolphins nach Belieben und gewannen vernichtend mit 20:3.

ES: Die Musiker spielten nach Belieben, bis der Dirigent auf das Podium trat.

Target Word # 7: knabbern (to nibble)

OS: Der Zuschauer im zweiten Stock des Weißen Hauses lag auf der Couch und knabberte an Brezeln.

ES: Die Kinder knabberten an ihren Keksen.

Target Word # 8: gesetzt den Fall (assuming that)

OS: Gesetzt den Fall, der Präsident lag auf seinem Sofa gegenüber dem Fernsehapparat, dann müßte er, sobald die ungekonnt verschlungene Brezel eine Ohnmacht auslöste, nach hinten gekippt, oder, wenn er denn nach vorne gekippt ist, in sich zusammengesunken sein.

ES: Du bist heute krank, aber gesetzt den Fall, Du bist morgen wieder gesund, kannst Du für mich einkaufen gehen.

Target Word # 9: Lageskizze (sketch-map)

OS: Eine genaue Lageskizze über Größe und Anordnung der Möbel im zweiten Stock blieben die Präsidentenhelfer zwar schuldig, sonst aber nichts.

ES: Die Journalisten zeichneten eine genaue Lageskizze über Größe und Anordnung der Möbel.

Target Word # 10: Kronzeuge (principal witness)

OS: Wie lange der Präsident aber wirklich besinnungslos lag, weiß niemand genau, außer Spot und Barney, die der Niedergesunkene als Kronzeugen für die kurze Verweildaue auf dem Teppich anführte:

ES: Ein Mitglied der Verbrecherbande verriet seine Kollegen und trat vor dem Gericht als Kronzeuge auf.

Target Word # 11: Verweildauer (retention period)

OS: Wie lange der Präsident aber wirklich besinnungslos lag, weiß niemand genau, außer Spot und Barney, die der Niedergesunkene als Kronzeugen für die kurze Verweildauer auf dem Teppich anführte:

ES: Die Verweildauer von flüssigen Speisen im Magen ist kürzer als die von harten Speisen.

Target Word # 12: Missgeschick (misfortune, mishap)

OS: Im gesundheitsbesessenen Amerika empfiehlt es sich im Übrigen, den Schwächeanfall des Präsidenten als Missgeschick mit einer Brezel darzustellen.

ES: Ihm passieren ständig Unfälle, und er hat im Leben immer Pech: er wird vom Missgeschick verfolgt.

Target Word # 13: verschwurbeln (to contort, screw up)

OS: Und seine Landsleute wissen ja zur Genüge, dass Bush zum Tölpeln neigt, wenn ihm auch momentan deutlich weniger verschwurbelte Sätze aus dem Mund fallen als vor dem 11.September.

ES: Der Betrunkene sprach in wirren, konfusen, sinnlosen, verschwurbelten Sätzen.

Target Word # 14: Unterling (subordinate)

OS: Als er ins Wochenende gegangen sei, so erzählten seine Unterlinge nach dem Zwischenfall beim Fernsehen, habe Bush geklagt, er fühle sich müde und mürbe.

ES: Der König befahl seinen Unterlingen, den Gefangenen hinzurichten.

Target Word # 15: mürbe (crumbly, soft, worn down)

OS: Als er ins Wochenende gegangen sei, so erzählten seine Unterlinge nach dem Zwischenfall beim Fernsehen, habe Bush geklagt, er fühle sich müde und mürbe.

ES: Das Leder seiner Jacke war alt, weich und mürbe.

Target Word # 16: Blessur (wound)

OS: Der gemeine Mensch neigt dann schon mal zu einem Schwächeanfall, der Blessuren nach sich zieht.

ES: Nach jedem Rugby-Spiel kommt er nach Hause und beschwert sich über kleine Schmerzen und Blessuren.

Target Word # 17: fressen (to eat, guzzle)

OS: Die meisten fressen ihren Groll still in sich hinein.

ES: Hunde fressen gerne rohes Fleisch.

Target Word # 18: Weltmacht (world power)

OS: Wer will in diesen Tagen der internationalen Einigkeit schon die einzige Weltmacht kritisieren.

ES: Amerika und Rußland sind zwei Weltmächte, die die Weltpolitik dominieren.

Target Word # 19: Anschlag (strike)

OS: Nach den Anschlägen am 11.September gehörten den Amerikanern die Sympathien der Welt - und Washington nutzte diesen Goodwill, appellierte bei der Zusammenstellung einer weltweiten Anti-Terror-Koalition auch an die hohen ethischen Ansprüche und Menschenrechtsgarantien einer freien Gesellschaft.

ES: Terroristen haben einen Anschlag auf den Politiker ausgeübt.

Target Word # 20: Einsatz (mission, deployment)

OS: Von einer Zivilisation, die es auch mit dem Einsatz militärischer Gewalt zu verteidigen gelte, war die Rede.

ES: Die Amerikaner haben einen militärischen Einsatz gegen Afghanistan angefangen.

Target Word # 21: beugen (to bend, knuckle down)

OS: Schnell beugte man sich in Berlin und anderswo zum Kniefall der "uneingeschränkten Solidarität":

ES: Der große Mann beugte sich, um mit der kleinen Frau zu sprechen.

Target Word # 22: fangen (to capture)

OS: Die amerikanische Regierung hat über hundert in Afghanistan gefangen genommene Taliban- und al-Qaida-Kämpfer von Kandahar nach Kuba verschleppt - ausgerechnet nach Guantanamo Bay, wo die US-Verfassung nicht gilt.

ES: Die Verbrecher sind gefangen worden und sind jetzt im Gefängnis.

Target Word # 23: lächerlich (ridiculous)

OS: Auf einen Militärstützpunkt, den die US-Regierung 1903 der karibischen Nation regelrecht abpresste (für den sie bis heute lächerliche 4085 Dollar jährliche "Pachtgebühr" bezahlt) und der als rechtliches Niemandsland gilt.

ES: Der Preis war lächerlich, weil er viel zu hoch war.

Target Word # 24: Internierte (internee)

OS: Washington hat die Internierten schon als die "schlimmsten Elemente" vorverurteilt, als "Männer, die Leitungen durchbeißen würden, um ein Flugzeug abstürzen zu lassen", aber noch nicht einmal Klage erhoben oder einen Rechtsbeistand erlaubt.

ES: Die Internierten waren in einem kleinen Gefängnis festgehalten.

Target Word # 25: Kampfhandlung (combat operation)

OS: Jeder, der bei Kampfhandlungen festgenommen wird, gilt als PoW und hat sofortigen und vollständigen Anspruch auf Schutz durch die Genfer Konvention (zumindest bis ein "kompetentes Tribunal" seinen endgültigen Status klärt).

ES: Jeder, der bei Kampfhandlungen in Afghanistan festgenommen wird, gilt als PoW.

Target Word # 26: rachsüchtig (vindictive)

OS: Zehn Jahre später traf Stern-Reporter Claus Lutterbeck eine rachsüchtige Ex-Geliebte in New Orleans.

ES: Sie war sehr rachsüchtig und versuchte bei jeder Gelegenheit, ihm das Böse, das er ihr angetan hatte, heimzuzahlen.

Target Word # 27: erniedrigen (to humiliate)

OS: Einerseits ist es erniedrigend, im Fernsehen als "falsche Blonde" vorgeführt zu werden.

ES: Juden wurden während der Zeit des Nationalsozialismus oft in der Öffentlichkeit zu erniedrigenden Handlungen gezwungen, zum Beispiel mussten in Wien manchmal Juden mit Zahnbürsten schmutzige Straßen putzen.

Target Word # 28: zücken (to pull out, produce)

OS: Begeistert zücken sie ihre Wegwerfkameras und rufen: "Gennifer, bitte noch ein Foto mit mir!"

ES: Bei jeder Sehenswürdigkeit zückte er seine Kamera und machte ein Foto.

Target Word # 29: vermeintlich (alleged)

OS: Denn als die vermeintliche Wahrheit ans Licht kam, ging die Hatz erst richtig los.

ES: Die vermeintliche Freundin von Klaus war in Wirklichkeit seine Schwester.

Target Word # 30: verübeln (to resent)

OS: Noch immer verübelt sie Clinton, dass er nicht offen zu der Affäre stand, sondern sie "verraten" hat.

ES: Er hat mir sehr verübelt, dass ich bei der Auseinandersetzung nicht für ihn Partei ergriffen habe.

Target Word # 31: schnaufen (to wheeze)

OS: Er joggte zu ihrem Apartmenthaus, schlüpfte durch die Hintertür in ihre Wohnung, vergnügte sich bei Gennifer und trabte dann schnaufend zurück zum Regierungssitz.

ES: Der Professor hatte beschlossen, im David Hume Tower die Treppen zu Fuß hinaufzugehen, und kam schließlich vor Ermüdung heftig schnaufend im 10.Stock an.

Target Word # 32: erwischen (to catch)

OS: Er war leichtsinnig, er hatte nie Angst, erwischt zu werden.

ES: Die Polizei hat die Einbrecher sofort nach der Tat erwischt.

Target Word # 33: Klage (lawsuit)

OS: Auf die Frage, ob sie die Geschichte nicht endlich ruhen lassen und ihre Klage gegen Hillary und deren Mitarbeiter wegen Rufschädigung zurückziehen könne, sagt Gennifer Flowers: “Ich will meinen Tag vor Gericht.”

ES: Er hat bei Gericht eine Klage eingereicht, da sein Nachbar mit seinen fast täglichen Partys schon über ein halbes Jahr lang die Nachtruhe nicht einhält.

Target Word # 34: hemmungslos (uninhibited)

OS: Mitarbeiter des Landes Niedersachsen surfen hemmungslos zu ihrem Privatvergnügen im Internet.

ES: Der Politiker hat seine Gegner in der anderen Partei hemmungslos angegriffen.

Target Word # 35: ledig (unmarried)

OS: Andreas B., 35, ledig, kinderlos, war zu Recht fristlos gefeuert worden, befanden die Richter.

ES: Ein Bruder von mir ist verheiratet, aber der andere ist noch ledig.

Target Word # 36: fristlos (without notice)

OS: Andreas B., 35, ledig, kinderlos, war zu Recht fristlos gefeuert worden, befanden die Richter.

ES: Andreas B. durfte keinen einzigen Tag im Job bleiben, sondern er wurde fristlos entlassen.

Target Word # 37: bislang (up to now)

OS: Bislang sind solche Urteile allerdings selten, räumt selbst der hannoversche Arbeitsrechtler Stefan Kramer ein, der den Fall für den Verband durchgefochten hat.

ES: Bislang durfte ich nicht Auto fahren; jetzt, da ich die Fahrprüfung bestanden habe, darf ich es.

Target Word # 38: EDV-Fachmann (computer scientist)

OS: Auch der Fall in Hannover flog nur auf, weil die Technik streikte und ein EDV-Fachmann den Computer inspizierte.

ES: Da mein Schwager EDV-Fachmann ist, hilft er mir, wenn mein Computer nicht funktioniert.

Target Word # 39: dröge (boring)

OS: Dabei ist privates Surfen längst zum Volkssport in den Büros geworden, und der deutsche Beamte, gelangweilt vom drögen Tun und Verrichten, macht sich seinen Arbeitstag offenbar besonders gern nett im Netz.

ES: Meine Schwester hat einen interessanten Job, aber mein Arbeitstag ist ziemlich dröge.

Target Word # 40: verrichten (to do one's job, perform)

OS: Dabei ist privates Surfen längst zum Volkssport in den Büros geworden, und der deutsche Beamte, gelangweilt vom drögen Tun und Verrichten, macht sich seinen Arbeitstag offenbar besonders gern nett im Netz.

ES: Er hat seine Arbeit jeden Tag pflichtbewußt verrichtet.

Target Word # 41: Rechner (computer)

OS: Wie hemmungslos Mitarbeiter im Öffentlichen Dienst ihre Rechner für ihr Freizeitvergnügen nutzen, belegt jetzt eine Untersuchung des Niedersächsischen Landesrechnungshofs.

ES: Da ich jetzt einen Rechner habe, kann ich die online-Ausgaben der Zeitungen lesen.

Target Word # 42: Landesrechnungshof (regional audit court)

OS: Wie hemmungslos Mitarbeiter im Öffentlichen Dienst ihre Rechner für ihr Freizeitvergnügen nutzen, belegt jetzt eine Untersuchung des Niedersächsischen Landesrechnungshofs.

ES: Der Landesrechnungshof hat geprüft, ob die Beamten im Ministerium ihre Aufgaben richtig erfüllen.

Target Word # 43: Landesbediensteter (regional civil servant)

OS: In der vergangenen Woche schickte die Behörde eine interne Mitteilung über die "Nutzung der Arbeitszeit von Landesbediensteten für private Internetrecherchen" an den niedersächsischen Ministerpräsidenten Sigmar Gabriel.

ES: Jeder Beamte ist gewissermassen ein Landesbediensteter.

Target Word # 44: beheben (to remedy, correct)

OS: Der Rechnungshof bat darin um Aufklärung, wie der durch "mangelnde Dienstaufsicht und Untätigkeit" verursachte "erhebliche Schaden" behoben werden könne.

ES: In meinem Job hatte ich früher viele Schwierigkeiten; jetzt sind sie aber gott sei Dank alle behoben.

Target Word # 45: Zugriff (access)

OS: Für ihre Studie werteten die Kontrolleure erstmals die Zugriffe von 20 000 Landesbediensteten aus, die sich über das Informatikzentrum Niedersachsen ins Internet einwählen.

ES: Je mehr Zugriffe meine Webseite aufweist, desto sicherer bin ich, dass die Leute sich für sie interessieren.

Target Word # 46: flanieren (to stroll)

OS: So flanieren Beamte besonders gern durch Online-Shops und bieten bei Internet-Auktionen mit.

ES: Am Wochenende flanierten sie die Promenade entlang, als ob sie die letzte Mode vorführten.

Target Word # 47: gravierend (serious)

OS: Obwohl schon Stellen abgebaut wurden, sei der Arbeitsdruck offenbar immer noch "nicht so gravierend, wie häufig von Dienststellenleitern oder Arbeitnehmervertretern herausgestellt wird."

ES: Der Richter sagte: "Das ist ein sehr gravierender Fall; ich verurteile Sie zu zwanzig Jahren Gefängnis."

Target Word # 48: Dienststellenleiter (chief of the office)

OS: Obwohl schon Stellen abgebaut wurden, sei der Arbeitsdruck offenbar immer noch "nicht so gravierend, wie häufig von Dienststellenleitern oder Arbeitnehmervertretern herausgestellt wird."

ES: Da ich morgen zum Arzt gehen soll, muß ich den Dienststellenleiter im Büro um einen freien Tag bitten.

Target Word # 49: Volkswirtschaft (national ecocomy, ecocomics)

OS: Allein 470 Millionen Dollar soll die Lewinsky-Affäre die amerikanische Volkswirtschaft gekostet haben, weil sich Arbeitnehmer den Starr-Report auf ihren Rechner luden, um über die sexuellen Vorlieben ihres damaligen Präsidenten Bill Clinton orientiert zu sein.

ES: Wenn jeder Arbeitnehmer ständig tüchtig arbeiten würde, würde unsere Volkswirtschaft viel gesünder.

Target Word # 50: traktieren (to maul)

OS: Das Informatikzentrum Niedersachsen will die recherchefreudigen Beamten jedenfalls nicht mit technischen Sperren traktieren.

ES: Gestern gab es überhaupt nichts; aber heute wurden wir mit Bier und Wurst traktiert.

Target Word # 51: Rahmendienstanweisung (work procdure)

OS: Schließlich gelte ja noch die Rahmendienstanweisung des Landesfinanzministeriums, nach der jeder Internet-Nutzer unterschreiben müsse, das Netz nur zu dienstlichen Zwecken zu nutzen.

ES: Nach der Rahmendienstanweisung für meinen Beruf muss ich in der Regel 40 Wochenstunden arbeiten.

Target Word # 52: Abmahnung (reprimand)

OS: Doch Kontrollen gibt es offenbar nicht; eine Abmahnung oder gar Kündigung erst recht nicht.

ES: Wenn sie einmal gegen diese Regel verstoßen, bekommen Sie vom Chef eine Abmahnung; passiert es ein zweites Mal, so werden Sie gefeuert.

Target Word # 53: Entlastung (relief)

OS: Möglicherweise aber ist durch Dauersurfen wenigstens an einer Stelle Entlastung für den Staatshaushalt zu erwarten - wer privat surft, hat schließlich weniger Zeit zum privaten Telefonieren.

ES: Die alte Frau sagte: “Wenn Sie meinen Koffer zum Bus tragen wollen, so ist das sicher eine Entlastung für mich.”

Target Word # 54: Stichprobe (random inspection, control sample)

OS: Nach einer Stichprobe aus dem Jahr 2000 wurden Dienstapparate des Landes Niedersachsen noch zu 30 Prozent privat genutzt.

ES: Es wird zwar nicht jeder Fall untersucht, aber wir müssen wenigstens bei einigen KollegInnen eine Stichprobe machen.

Target Word # 55: verderben (to addle, decay)

OS: Sie kamen in Sturmfluten und Feuersbrünsten um, sie verhungerten, weil die Ernte auf den Feldern verdarb, oder sie fielen Seuchen zum Opfer, die sich im Gefolge des Wetterdurcheinanders ausbreiteten.

ES: Das Obst verdarb sehr schnell, da es schon beim Kauf nicht mehr frisch war.

Target Word # 56: Seuche (epidemic, plague)

OS: Sie kamen in Sturmfluten und Feuersbrünsten um, sie verhungerten, weil die Ernte auf den Feldern verdarb, oder sie fielen Seuchen zum Opfer, die sich im Gefolge des Wetterdurcheinanders ausbreiteten.

ES: Im Mittelalter war die Pest eine verbreitete Seuche.

Target Word # 57: Voraussetzung (precondition)

OS: "Die wesentlichen Voraussetzungen dafür sind gegeben", warnte soeben US-Meteorologe Stephen Zebiak auf einem Wissenschaftlerkongress.

ES: Das Abitur ist eine Voraussetzung für ein Universitätsstudium.

Target Word # 58: Öllache (oil slick)

OS: Ein zunehmend breiter werdender Warmwasserteppich - Vorbote drohenden Unheils - schwappt träge wie eine Öllache ostwärts.

ES: Als wir die Öllache sahen, wussten wir, dass das Auto Öl verliert.

Target Word # 59: Überschwemmung (inundation)

OS: Überschwemmungen und Hagelstürme drohen den süd- und nordamerikanischen Westküsten.

ES: Bei den Überschwemmungen in Südengland hatten viele Leute tagelang Wasser im Keller.

Target Word # 60: Dürre (drought)

OS: Schwere Dürren, die die gewohnten klimatischen Bedingungen auf den Kopf stellen, werden in Australien, Neuseeland und in weiten Teilen Südostasiens erwartet.

ES: Die Dürre in Afrika verursachte eine große Hungersnot.

Target Word # 61: nachweisbar (detectable)

OS: Rätselhaft ist, warum die seit mindestens 130 000 Jahren nachweisbare Unwetter-Konstellation neuerdings immer häufiger und stärker über den Planeten hereinbricht.

ES: Er hatte Glück, denn im Blut war kein Alkohol nachweisbar.

Target Word # 62: Zeche (check)

OS: Sollte El Niño die Wetterküche auch in diesem Jahr aufwühlen, würden erneut vor allem die Armen der Welt die Zeche bezahlen.

ES: Sie verließen die Bar ohne die Zeche zu bezahlen.

Target Word # 63: Schaden (damage)

OS: Die Kapriolen von 1997/98 haben weltweit Schäden in Höhe von weit über 30 Milliarden Dollar verursacht.

ES: Der Sturm verursachte Schäden in Millionenhöhe.

Target Word # 64: zerstörerisch (destructive)

OS: Bei manchen kommt es immer wieder - regelmäßig und zerstörerisch.

ES: Fußballfans sind oft zerstörerisch nachdem sie verloren haben: sie machen dann Dinge kaputt.

Target Word # 65: bedrückend (depressing, burdensome)

OS: Aber die Niederlage ist dramatisch und bedrückend.

ES: Die Hitze ist bedrückend.

Target Word # 66: unberechenbar (incalculable)

OS: Warum sind die Ossis als Wahlvolk so unberechenbar?

ES: Das schottische Wetter ist unberechenbar, mal regnet es, mal schneit es.

Target Word # 67: Ossi (East German)

OS: Warum sind die Ossis als Wahlvolk so unberechenbar?

ES: Die Ossis aus Leipzig stritten sich mit den Wessis aus Stuttgart.

Target Word # 68: Enttäuschbarkeit ('ease of being disappointed')

OS: Das hat vor allem mit der leichten Enttäuschbarkeit zu tun.

ES: Ihre Enttäuschbarkeit äußerte sich in häufigen Tränenausbrüchen.

Target Word # 69: versäumen (to neglect)

OS: In Sachsen-Anhalt wurde wohl auch versäumt, frühzeitig die Grenzen staatlicher Handlungsmöglichkeiten aufzuzeigen und den Menschen zu sagen, dass sie noch mehr tun müssen als bisher schon.

ES: Ich habe versäumt, ihn anzurufen, deswegen ist er nun sauer.

Target Word # 70: Handlungsmöglichkeit (option to act)

OS: In Sachsen-Anhalt wurde wohl auch versäumt, frühzeitig die Grenzen staatlicher Handlungsmöglichkeiten aufzuzeigen und den Menschen zu sagen, dass sie noch mehr tun müssen als bisher schon.

ES: Meine Handlungsmöglichkeiten sind begrenzt, ich habe einfach nicht genug Geld.

Target Word # 71: unerfreulich (unpleasant)

OS: Das darf man nicht überbewerten, das ist nur für die SPD unerfreulich.

ES: Ich habe eine Grippe bekommen, das ist sehr unerfreulich.

Target Word # 72: Freiraum (free space, leeway)

OS: Obwohl die PDS in Magdeburg - halb regierte sie mit, halb blieb sie Opposition - viel Freiraum hatte, hat sie nichts dazugewonnen, sondern ihre stabile Stammwählerschaft mobilisiert.

ES: Mein Beruf läßt mir viel Freiraum, da ich jeden Abend früh nach Hause kann.

Target Word # 73: Stammwählerschaft (group of loyal voters)

OS: Obwohl die PDS in Magdeburg - halb regierte sie mit, halb blieb sie Opposition - viel Freiraum hatte, hat sie nichts dazugewonnen, sondern ihre stabile Stammwählerschaft mobilisiert.

ES: Mein Vater gehört zur Stammwählerschaft der CDU.

Target Word # 74: befremdlich (strange)

OS: Aus westlicher Sicht und auch aus Sicht von DDR-Oppositionellen ist das trotzdem etwas Befremdliches: dass die Ostdeutschen nun die Nachfolger jener Partei wählen, die sie einst in den Ruin gebracht hat.

ES: Er sieht so seltsam und befremdlich aus.

Target Word # 75: gegenwärtig (current)

OS: Aber man muss auch daran erinnern, dass man die gegenwärtigen materiellen und emotionalen Verhältnisse im Osten nicht mehr allein mit der Erbschaft aus DDR-Zeiten erklären kann.

ES: Die gegenwärtige Situation in Schottland ist entspannt, es ist sehr friedlich heute.

Target Word # 76: labil (weak, unstable)

OS: Die Ostdeutschen haben erlebt, wie labil ein politisches System sein kann.

ES: Meine Großmutter ist schon 80 Jahre alt und ziemlich labil.

Target Word # 77: Hornhaut (horny skin)

OS: Die Ostdeutschen haben noch nicht die demokratische Hornhaut, um angesichts der Fehlleistungen und Enttäuschungen in der Politik nicht gleich das ganze System anzuzweifeln.

ES: Er ist immer viel ohne Schuhe herumgelaufen, daher hat er Hornhaut an den Füßen.

Target Word # 78: latent (latent)

OS: Das ist eine gefährliche Stimmung, weil sie latent antidemokratische Unsicherheiten überspielt und diese zu antidemokratischen Vorurteilen verfestigt.

ES: Er ist latent schwul, hat sich das aber noch nicht eingestanden.

Target Word # 79: verfestigen (to solidify)

OS: Das ist eine gefährliche Stimmung, weil sie latent antidemokratische Unsicherheiten überspielt und diese zu antidemokratischen Vorurteilen verfestigt.

ES: Der Zement hat sich verfestigt.

Target Word # 80: schäbig (mean, seedy, run-down)

OS: Der ist ja genauso schäbig, wie sie uns oder unseren Herrschaften immer vorgeworfen haben.

ES: Das schäbige Haus steht im Ghetto der Stadt.

Target Word # 81: Nachholbedarf (backlog demand)

OS: Ich glaube, da haben Ostdeutsche wirklich noch einen erheblichen Nachholbedarf.

ES: Ich lag eine ganze Woche krank im Bett und war nicht draußen, deswegen habe ich nun einen grossen Nachholbedarf.

Target Word # 82: Ungleichgewicht (imbalance)

OS: Dadurch ergab sich ein Ungleichgewicht und eben keine Gleichberechtigung.

ES: Er ist viel schwerer als sie, daher herrscht zwischen den beiden ein Ungleichgewicht.

Target Word # 83: Gleichberechtigung (equality)

OS: Dadurch ergab sich ein Ungleichgewicht und eben keine Gleichberechtigung.

ES: Es herrscht keine Gleichberechtigung zwischen Ost- und Westdeutschland.

Target Word # 84: Minderwertigkeitsgefühl (sense of inferiority)

OS: Das setzt natürlich die schlechte Tradition des ostdeutschen Minderwertigkeitsgefühls fort:

ES: Er ist sehr schüchtern und hat scheinbar ein großes Minderwertigkeitsgefühl.

Target Word # 85: eisern (steely, iron-clad)

OS: Blickt nicht immer nur eisern nach Westen.

ES: Er starrt eisern auf das Bild, sein Blick bewegt sich nicht.

Target Word # 86: meistern (to master)

OS: Jetzt müssen wir auf uns selber gucken und auf das, was wir in den vergangenen zwölf Jahren gemeistert haben in dieser gigantischen Transformation.

ES: Er hat die Klausur spielerisch gemeistert und eine sehr gute Note bekommen.

Target Word # 87: jammern (to moan, lament)

OS: Die jammern gern auf höherem Niveau.

ES: Die Katze jammert an der Tür, sie will ins Haus.

Target Word # 88: Habitus (disposition)

OS: Die Öffentlichkeit wird von den Medien bestimmt, und die Medien in Deutschland sind westlich, personell wie sprachlich, in ihrem Habitus, in Stil, Wahrnehmungen, Schwerpunktsetzungen.

ES: Der Habitus des Affen ist sehr charakteristisch.

Target Word # 89: Wahrnehmung (perception)

OS: Die Öffentlichkeit wird von den Medien bestimmt, und die Medien in Deutschland sind westlich, personell wie sprachlich, in ihrem Habitus, in Stil, Wahrnehmungen, Schwerpunktsetzungen.

ES: Seine Wahrnehmung ist getrübt, er denkt dass Frankensteins Monster hübsch ist.

Target Word # 90: Schwerpunktsetzung (emphasis)

OS: Die Öffentlichkeit wird von den Medien bestimmt, und die Medien in Deutschland sind westlich, personell wie sprachlich, in ihrem Habitus, in Stil, Wahrnehmungen, Schwerpunktsetzungen.

ES: Die Schwerpunktsetzung bei dem Germanistik-Studium in Glasgow liegt bei der Romantik.

Target Word # 91: Grundton (keynote, 'basic tint')

OS: Ich glaube, es fängt schon mit einem bestimmten Grundton an, dem Grundton der Häme.

ES: Sein Grundton ist pessimistisch, er sieht ständig schwarz.

Target Word # 92: Häme (malice)

OS: Ich glaube, es fängt schon mit einem bestimmten Grundton an, dem Grundton der Häme.

ES: Er war schadenfroh wie immer und betrachtete ihr gebrochenes Bein voller Häme.

Target Word # 93: Feldstecher (field-glasses)

OS: Ausgerüstet mit Feldstechern, suchten sie die Hänge und Felder rund um das Absturzgebiet der DHL-Frachtmaschine ab.

ES: An den Küsten Schottlands kann man viele Vogelarten mit Hilfe von Feldstechern beobachten.

Target Word # 94: schwappen (to swash, spill)

OS: Das war beim Oder-Hochwasser 1997 so, und auch die Kölner Altstadtbewohner kennen diese durchaus unwillkommenen Besucher, die so regelmäßig in der Domstadt erscheinen, wie der Rhein über die Ufer schwappt.

ES: Wenn der Wasserspiegel des Rheins steigt, schwappt er manchmal über die Ufer und überflutet Köln.

Target Word # 95: Nervenkitzel (thrill)

OS: "Die brauchen das als Nervenkitzel", versucht sich ein Göppinger Bereitschaftspolizist in Tiefenpsychologie, bevor er die lästigen Gaffer vertreibt, die sich nun einen anderen Ausguck suchen müssen.

ES: Viele Extremsportarten sind so beliebt, weil sie einen gewissen Nervenkitzel bieten.

Target Word # 96: Gaffer (gaper)

OS: "Die brauchen das als Nervenkitzel", versucht sich ein Göppinger Bereitschaftspolizist in Tiefenpsychologie, bevor er die lästigen Gaffer vertreibt, die sich nun einen anderen Ausguck suchen müssen.

ES: Auf der Autobahn sind zwei Autos verunglückt, und jetzt sehen sich zahlreiche Gaffer die Szene an.

Target Word # 97: Ausguck (lookout)

OS: "Die brauchen das als Nervenkitzel", versucht sich ein Göppinger Bereitschaftspolizist in Tiefenpsychologie, bevor er die lästigen Gaffer vertreibt, die sich nun einen anderen Ausguck suchen müssen.

ES: In Burgen gibt es viele kleine Löcher in der Mauer, die häufig als Ausguck für die Burgbewohner dienen.

Target Word # 98: bergen (salvage)

OS: Tornados fliegen über den Absturzkorridor und machen Luftaufnahmen, damit alle Opfer so schnell wie möglich gefunden und geborgen werden können.

ES: Nach dem Schiffsunglück versuchen nun Rettungsteams, die Opfer zu bergen.

Target Word # 99: haarscharf (by a whisker)

OS: Die Stadt Überlingen ist haarscharf einer Katastrophe entkommen.

ES: Die Katastrophe ist haarscharf an uns vorbeigegangen, beinahe wären viele Menschen ums Leben gekommen.

Target Word # 100: Trümmer (wreckage)

OS: Die Trümmer der beiden Maschinen schlugen an insgesamt 57 Stellen in einem Umkreis von 30 Kilometern auf, ohne größere Schäden anzurichten.

ES: Nach der Bombenexplosion lag das Haus in Trümmern.

Target Word # 101: Tragflächenteil (wing part)

OS: In Owingen krachten Fahrwerk und Tragflächenteile in einen Garten und verfehlten knapp ein Wohnhaus.

ES: Die Tragflächen sind zerbrochen, und deshalb liegen viele Tragflächenteile auf dem Boden.

Target Word # 102: Leuchtpistole (signal pistol)

OS: "Sind die denn verrückt geworden, mitten in der Nacht noch Leuchtpistolen abzuschießen?", dachte sich Hermann Schmidt, der zusammen mit seiner Familie auf einem idyllisch gelegenen Bauernhof Ferien machte.

ES: Ein Schiff in Seenot schiesst mit Leuchtpistolen, um Flugzeuge und andere Schiffe auf sich aufmerksam zu machen.

Target Word # 103: Sachlage (circumstance)

OS: Doch dieses Mal erscheint die Sachlage nicht so eindeutig erklärbar.

ES: Die Sachlage ist die: es gab einen Unfall, den der Fahrer des blauen Autos verschuldet hat.

Target Word # 104: Steuerknüppel (yoke, joystick)

OS: Indem der baskirische Pilot den Steuerknüppel nach vorne drückte, um endlich den von der Flugsicherung verlangten Sinkflug einzuleiten, war die Katastrophe nicht mehr aufzuhalten.

ES: Der Pilot fliegt das Flugzeug mit dem Steuerknüppel, den er in der Hand hält.

Target Word # 105: orten (to locate)

OS: In jenem Moment hatte nämlich das automatische Warnsystem der Boeing 757 das fliegende Hindernis geortet und dem Piloten ebenfalls den Sinkflug befohlen.

ES: Die Flugsicherung ortet alle Flugzeuge im Luftraum.

Target Word # 106: unverzüglich (immediately)

OS: Diesem Befehl, das entspricht der international geltenden Absprache, muss der Pilot unverzüglich und ohne weitere Rücksprache mit der Flugsicherung Folge leisten.

ES: Der Pilot muss unverzüglich, also sofort, auf die Flugsicherung hören.

Target Word # 107: zügig (speedy)

OS: So fragt auch die Pilotenvereinigung Cockpit, ob die beteiligten Flugsicherungen in Deutschland, Österreich und der Schweiz tatsächlich alle Informationen über die beiden Flüge richtig und zügig weitergeleitet hätten.

ES: Es ist wichtig, Informationen zügig, also so schnell wie möglich, weiterzuleiten.

Target Word # 108: angewiesen (to rely on)

OS: Und dann kann die ganze Kollisionswarntechnik nicht funktionieren, sind die Piloten allein auf die Bodenkontrolle angewiesen.

ES: Die Piloten müssen der Bodenkontrolle vertrauen, weil sie ganz auf sie angewiesen sind.

Target Word # 109: Ursprungsland (country of origin)

OS: In Großbritannien, dem Ursprungsland der BSE-Seuche, gibt man dem Kontinent die Schuld an der neuen Misere.

ES: Obwohl Argentinien das Ursprungsland des Tangos ist, ist dieser Tanz auch in Europa bekannt.

Target Word # 110: Ansteckungsgefahr (risk of infection)

OS: Kein Mensch brauche sich, hieß es, über eine Ansteckungsgefahr Gedanken zu machen:

ES: Wegen der Ansteckungsgefahr musste er 4 Wochen in Quarantäne.

Target Word # 111: schmählich (ignominious)

OS: Schockiert aber waren doch viele Briten, ihr Wappentier und andere Großkatzen der Nation so schmählich in die Knie sinken zu sehen.

ES: Es war ein schmählicher Trost, dass sie statt Letzte nur Vorletzte geworden war.

Target Word # 112: unverblümt (blunt)

OS: "Ich gehe schon davon aus", erklärte unverblümt Newquays Zoodirektor Mike Thomas, "dass die Erkrankung davon herrührte, dass der Löwe Stücke ganzer Kadaver verzehrt hat - wir wissen ja, dass die Krankheit vor allem über Gehirn und Rückgrat übertragen wird."

ES: Ich war ziemlich schockiert, als er mir unverblümt sagte, ich sei viel zu dick.

Target Word # 113: verzehren (to consume)

OS: "Ich gehe schon davon aus", erklärte unverblümt Newquays Zoodirektor Mike Thomas, "dass die Erkrankung davon herrührte, dass der Löwe Stücke ganzer Kadaver verzehrt hat - wir wissen ja, dass die Krankheit vor allem über Gehirn und Rückgrat übertragen wird."

ES: Sie hatte noch keinen Bissen verzehrt, als der Kellner den Tisch schon wieder abräumte.

Target Word # 114: zollen (to pay something *[fig.]*)

OS: Mit ihren verseuchten Löwen und Tigern, mit mehr als 177 000 Rindern und Kühen, denen BSE in den vergangenen zwölf Jahren zum Verhängnis wurde, und mit inzwischen über 80 Menschen, die vCJK, die menschliche Variante des Rinderwahns, das Leben kostete, haben die Bewohner des Vereinigten Königreichs der BSE-Epidemie einen hohen Preis gezollt.

ES: Angesichts der vielen Todesfälle bei Mensch und Tier haben die Briten der Epidemie einen hohen Preis gezollt.

Target Word # 115: Früherkennung (early diagnosis)

OS: Die Beef-Seuche, die so lange keine sein durfte, und deren wahres Ausmaß immer noch niemand abzuschätzen vermag, hat mittlerweile dazu geführt, dass die Regierung Tests an Lämmern verordnet hat und im Notfall die gesamten Schafbestände Britanniens schlachten lassen will, und dass britische Krankenhäuser sich ernsthaft um die Blutreserven des Landes Sorgen machen, da bisher keine Tests zur Früherkennung von vCJK verfügbar sind.

ES: Die Mammografie ist zur Früherkennung von Brustkrebs unerläßlich und kann das Leben vieler Frauen retten.

Target Word # 116: überschreiten (to cross)

OS: Andererseits sind die Briten davon überzeugt, dass der Höhepunkt der BSE-Epidemie auf der Insel überschritten ist - auch wenn noch immer tausend Rinder im Jahr neu diagnostiziert werden, und die Konsequenzen der Krise, wegen der langen Inkubationszeit, erst jetzt voll durchschlagen.

ES: Als der illegale Einwanderer nachts die Grenze überschreitet, wird er auf der anderen Seite von Scheinwerferlicht begrüßt.

Target Word # 117: rückläufig (declining)

OS: Die BSE-Zahlen indes sind schon seit einiger Zeit rückläufig.

ES: Früher erkrankten viele Menschen an Tuberkulose, zum Glück sind die Zahlen seit kurzem rückläufig.

Target Word # 118: wacker (courageous)

OS: Während Franzosen und Deutsche in ihrer aktuellen Panik Rindfleisch von der Speisekarte streichen, greifen die Briten in ihren Metzgereien und Supermärkten wieder wacker zu.

ES: Ohne Angst ging er auf den Stier zu und packte ihn wacker bei den Hörnern.

Target Word # 119: Operationsbesteck (surgical instruments)

OS: Der weit verbreitete Glaube, fleischmäßig “überm Berg” zu sein, führt heute zu der paradoxen Situation, dass immer mehr Briten sich wieder ihr Beefsteak schmecken lassen, während die Zahl der Creutzfeldt-Jakob-Toten auf der Insel wächst und Englands Ärzte sich darüber streiten, ob sie nach Zahn- und Blinddarm-Operationen das Operations-Besteck wegwerfen sollen oder nicht - vCJK-Erreger lassen sich mit normalen Desinfektions-Prozeduren nicht so leicht vernichten wie Bakterien oder Viren.

ES: Nach jeder Operation werden die Operationsbestecke sterilisiert.

Target Word # 120: verbannen (to ban, exile)

OS: Auch an vielen britischen Schulen, die in den 90er Jahren Rindfleisch aus ihren Küchen verbannten, taucht es mittlerweile wieder auf der Speisekarte auf.

ES: Als Napoleon verbannt wurde, ging er nach Elba ins Exil.

Target Word # 121: Abschottung (sealing-off)

OS: Selbstbewusst verlangen neuerdings Beef-Produzenten und andere Nutznießer des Fleischgewerbes im Königreich Abschottung des heimischen Marktes gegen "gefährliche Importe" vom Kontinent:

ES: Trotz der perfekten Abschottung griff der Feind die Burg an.

Target Word # 122: genüßlich (with pleasure)

OS: Geradezu genüßlich haben konservative Politiker und anti-europäische Zeitungen der Insel auf die neue Misere in Frankreich und auf den teutonischen Schock über das Ende der Selbsttäuschung in Deutschland reagiert.

ES: Genüßlich biss sie in die Schokolade, statt sie auf einmal hinunterzuschlingen.

Target Word # 123: Selbsttäuschung (self-deception)

OS: Geradezu genüßlich haben konservative Politiker und anti-europäische Zeitungen der Insel auf die neue Misere in Frankreich und auf den teutonischen Schock über das Ende der Selbsttäuschung in Deutschland reagiert.

ES: Seine Lügen führten zur Selbsttäuschung, so dass er letztendlich selbst nicht mehr Wahrheit von Lüge unterscheiden konnte.

Target Word # 124: befugen (to authorize)

OS: Nach vier Jahren kleinlauter Töne fühlt die britische Rechte sich wieder zum Aufdrehen des Volumens befugt.

ES: Obwohl er dazu nicht befugt war, nahm er sich selbst den Schlüssel, ohne auf den Pförtner zu warten.

Target Word # 125: Ursprung (origin)

OS: Die Ursprünge der Krise, im Vormonat erst auf 4000 Seiten von einem Untersuchungs-Ausschuß vor der Öffentlichkeit ausgebreitet, bemüht man sich geflissentlich zu vergessen.

ES: Der Rhein hat seinen Ursprung in der Schweiz.

Target Word # 126: geflissentlich (assiduous)

OS: Die Ursprünge der Krise, im Vormonat erst auf 4000 Seiten von einem Untersuchungs-Ausschuß vor der Öffentlichkeit ausgebreitet, bemüht man sich geflissentlich zu vergessen.

ES: Geflissentlich sah sie darüber hinweg, dass er schon wieder ihren Hochzeitstag vergessen hatte.

Target Word # 127: Unverschämtheit (impertinence)

OS: Als "kolossale Unverschämtheit" charakterisiert Lichfield den gegenwärtigen Versuch des konservativen Lagers seines Landes, "anti-europäisches Kapital" aus der Tatsache zu schlagen, dass nun auch in kontinentalen Rinderherden BSE entdeckt worden sei.

ES: "Was für eine Unverschämtheit", dachte die alte Dame, als sich der Junge auf den letzten freien Platz setzte, obwohl auch eine hochschwangere Frau mit eingestiegen war.

Target Word # 128: zugetan (affectionate, to like)

OS: Auch die Regierung von Tony Blair, "patriotischen" Gesten weniger zugetan als Tory-Regierungen der Vergangenheit, sucht einstweilen kühlen Kopf zu wahren, und hat sich bisher geweigert, ein Embargo gegen ausländisches Beef zu verhängen, wie es von der Opposition gefordert wird.

ES: Dass er dem Mädchen sehr zugetan war, sah man daran, dass er ihr stets hilfreich zur Seite stand.

Target Word # 129: kreuzbrav (very well-behaved)

OS: Der tritt als kreuzbraver Europäer auf, es sei denn, er spricht vor den Vertriebenen.

ES: Das ist ein kreuzbraver Mensch, der so etwas tut.

Target Word # 130: Augenblinzeln ('eyelash-fluttering')

OS: Und wie kann man da populär werden, ohne einerseits als Entertainer zu glänzen und andererseits mit Augenblinzeln und Andeutungen - also ohne das Vulgäre von Westerwelle & Möllemann - auch jenen ein Obdach zu bieten, die sich in den grossen Parteien derzeit nicht mehr zu Hause fühlen?

ES: Die Frau kommunizierte mit dem attraktiven Mann in der Bar mit Augenblinzeln.

Target Word # 131: schummrig (dim)

OS: In dem grossen schummrigen Kommunikationsraum, in dem wir uns alle bewegen, in dem aber die klassischen Rollen von Politik, Medien, Demoskopie, Werbung, Wissenschaft und Sachverstand sich ineinander aufgelöst haben, erscheint der Populismus nur plötzlich als das klar erkennbare Vis-a-vis.

ES: Diese Kneipe ist bei Liebespaaren beliebt, denn sie hat viele schummrige Ecken.

Target Word # 132: rumoren (to brew *[fig.]*)

OS: Es rumort gefährlich.

ES: Im Dampfkessel rumort es, er explodiert bald.

Target Word # 133: angeschlagen (groggy)

OS: Als die Abgeordneten wenig später - in der ersten Sondersitzung seit neun Jahren - die Risiken der Afghanistan-Mission und einen möglichen Angriff auf den Irak diskutierten, war der angeschlagene Premier schon wieder weg.

ES: Der Boxer war schon angeschlagen, aber noch nicht k.o.

Target Word # 134: wacklig (unfirm)

OS: Seit sich der Premier dem Krieg gegen den Terrorismus verschrieben hat, war die Heimatfront noch nie so wacklig.

ES: Der Betrunkene steht auf wackligen Beinen.

Target Word # 135: gären (to ferment)

OS: Nachdem bereits 131 Labour-Abgeordnete einen Angriff auf den Irak verurteilt hatten, gärt jetzt selbst im Kabinett der Widerstand.

ES: Wenn Traubensaft gärt, entsteht Alkohol und Wein.

Target Word # 136: Neuauflage (new edition)

OS: In einer Meinungsumfrage sprachen sich lediglich 35 Prozent der Befragten für eine britische Beteiligung an einer Neuauflage des Golfkriegs aus.

ES: Das Buch ist ausverkauft, aber weil es so erfolgreich war, plant der Verlag eine Neuauflage.

Target Word # 137: Feldzug (campaign)

OS: Auf dem letzten EU-Gipfel in Barcelona versuchte der Brite vergebens, europäische Amtskollegen für einen Feldzug gegen Saddam zu rekrutieren.

ES: Die Armee brach zu einem Feldzug gegen das Nachbarland auf.

Target Word # 138: nachhaltig (lasting)

OS: Gleichzeitig brachte er die Gewerkschaften daheim nachhaltig gegen sich auf, indem er mit Silvio Berlusconi eine Allianz gegen Frankreich und Deutschland begründete, um die Arbeitnehmerrechte in der EU zu schwächen.

ES: Die Maßnahmen sollen nicht nur kurzfristig wirken, sondern nachhaltig sein.

Target Word # 139: anrüchig (dubious, objectionable)

OS: Anrüchige Parteispenden und Nepotismus am Hof des Premiers, der wie ein amerikanischer Präsident regiert, haben das Vertrauen zermürbt.

ES: Diese Angelegenheit stinkt, mit anrüchigen Geschäften will ich nichts zu tun haben.

Target Word # 140: Hof (court)

OS: Anrüchige Parteispenden und Nepotismus am Hof des Premiers, der wie ein amerikanischer Präsident regiert, haben das Vertrauen zermürbt.

ES: Als der Prinz kam, schliefen am Hof von Dornröschens Vater alle.

Target Word # 141: zermürben (to wear down)

OS: Anrüchige Parteispenden und Nepotismus am Hof des Premiers, der wie ein amerikanischer Präsident regiert, haben das Vertrauen zermürbt.

ES: Eine Foltermethode bestand darin, Gefangene zu zermürben, indem man ihnen tagelang Wasser auf den Kopf tropfen ließ.

Target Word # 142: dröge (boring)

OS: Innerhalb eines Monats schrumpfte der Vorsprung, den Labour in Umfragen vor den drögen Konservativen hat, von 17 auf 7 Prozent.

ES: Niemand hat ihn je lachen sehen, das ist wirklich ein dröger Mensch.

Target Word # 143: karikieren (to caricature)

OS: Selbst die “Times” karikiert Blair mittlerweile als Pudel mit der US-Flagge als Fell.

ES: In dieser Zeichnung ist die Person nicht naturgetreu dargestellt, sondern übertrieben und karikiert.

Target Word # 144: preisen (to praise)

OS: So wird Blair zwar in Amerika als der treueste Verbündete gepriesen, wenn er den Kreuzzug gegen das Böse predigt.

ES: Viele Briten preisen Churchill als den besten Premierminister dieses Jahrhunderts.

Target Word # 145: verwinden (to get over)

OS: Brown ist in der Partei beliebt und hat bis heute nicht verwunden, dass er 1994 beim Kampf um den Labour-Vorsitz gegen Blair den Kürzeren zog.

ES: Der Witwer trauert noch immer, denn er hat den Tod seiner Frau nicht verwunden.

Target Word # 146: Mißtrauensantrag (motion of no confidence)

OS: Der Labour-Mann aus Schottland räumt zwar ein, dass die Wahrscheinlichkeit eines Mißtrauensantrags gegen Blair gering ist.

ES: In Deutschland kann die Opposition nur dann einen Mißtrauensantrag gewinnen, wenn sie einen neuen Kanzler wählen kann.

Target Word # 147: entpuppen (to turn out to be something/someone)

OS: Hat sich der "spannendste Job" nicht längst als Höllenkommando entpuppt?

ES: Aus seiner Idee ist nichts geworden, und sie hat sich als völlige Farce entpuppt.

Target Word # 148: entschärfen (to defuse)

OS: Tage später steckte meine Frau Doris in einem künstlich erzeugten Autostau, in dem in letzter Sekunde eine Bombe entschärft wurde.

ES: Die Bombe wurde rechtzeitig entschärft - fünf Minuten später wäre sie explodiert.

Target Word # 149: Gradmesser (indicator)

OS: Sie ist wohl der sensibelste Gradmesser für die politische Situation.

ES: Die guten Schulnoten beweisen die hohe Qualität des Unterrichts: Sie sind ein Gradmesser, also ein klarer Indikator dafür.

Target Word # 150: Leibesvisitation (body search)

OS: Wir haben auch keinerlei Probleme, dass ohne Leibesvisitation überhaupt nichts geht, auch der Kofferraum jedesmal schärfstens kontrolliert wird.

ES: Bei Polizeikontrollen am Flughafen werden Leibesvisitationen gemacht und die Fluggäste untersucht, ob sie z.B. Waffen am Körper tragen.

Target Word # 151: Gräueltat (atrocity)

OS: Er hat vor der Diskothek Blumen niedergelegt, ist anschließend zu Arafat gefahren und hat ihn erfolgreich bewogen, sich öffentlich von der abscheulichen Gräueltat zu distanzieren.

ES: Im Krieg wurden viele brutale Morde und andere Gräueltaten verübt.

Target Word # 152: versäumen (to miss [an opportunity])

OS: Selbst eine jordanische Zeitung titelte, dass Arafat damit die Chance seines Lebens versäumt hat.

ES: Die Studentin war immer anwesend und versäumte keine einzige Stunde.

Target Word # 153: Getümmel (turmoil)

OS: Da ich nicht mehr im Getümmel stecke, richte ich mich nach den Meinungsumfragen.

ES: Beim Sommerschlussverkauf waren so viele Menschen, dass ich in dem Getümmel meine Freundin verlor.

Target Word # 154: Schaltung (gear shift)

OS: Nicht die Handbremse und nicht die Schaltung.

ES: Wenn das Auto schneller wird, muss ich den Gang wechseln, also die Schaltung bedienen.

Target Word # 155: lichten (to clear, lift)

OS: Der Anker wird gelichtet.

ES: Der Wald lichtet sich.

Target Word # 156: anschmiegsam (soft, cuddly)

OS: Von außen wirkt er mächtig und knallhart, aber dank seiner Rundungen und dem weichen Dach zugleich sanft und anschmiegsam.

ES: Dieser neue, unglaublich weiche Pullover ist absolut anschmiegsam auf der Haut.

Target Word # 157: Vollblut (thoroughbred)

OS: Ein Vollblut, schwarz leuchtend.

ES: Dieses temperamentvolle Pferd ist ein echtes Vollblut.

Target Word # 158: wickeln (to wind, wrap)

OS: Zum ersten Mal verstehe ich, warum manche Männer ihre Autos so liebevoll waschen, wickeln und streicheln.

ES: Das Baby muss gewickelt werden.

Target Word # 159: Verdeck (canopy top, convertible top)

OS: Das Verdeck muss in der Parkposition beziehungsweise im Leergang abgetakelt werden.

ES: Unter dem Verdeck fanden wir Schutz.

Target Word # 160: Leergang (neutral *[mot.]*)

OS: Das Verdeck muss in der Parkposition beziehungsweise im Leergang abgetakelt werden.

ES: Es wird empfohlen, an der roten Ampel den Leergang einzulegen.

Target Word # 161: Markise (awning)

OS: Als hätte man an einem Sommermorgen die Jalousien hochgezogen oder die Markise zurückgerollt.

ES: Als die Gäste im Straßencafe die ersten Regentropfen spüren, rollt der Wirt die Markise über ihren Köpfen aus.

Target Word # 162: überriechen (to ignore a smell)

OS: Dinge, die man überhören, übersehen, überriechen muss.

ES: Den unangenehmen Geruch müsst ihr einfach "überriechen".

Target Word # 163: Zellenfahrer (convertible driver)

OS: Freunde haben mir ein Publikum Spalier stehender, einsteigewilliger Damen sowie neidisch blickender, eingesperrter Zellenfahrer in Aussicht gestellt.

ES: Cabriofahrer verachten alle "Zellenfahrer".

Target Word # 164: Abschleppeffekt ('tow-away effect')

OS: Keine Spur von Abschleppeffekt.

ES: Gutaussehende berühmte Persönlichkeiten wie Brad Pitt haben oft einen Abschleppeffekt auf Frauen.

Target Word # 165: zockeln (to trundle)

OS: Selbst wenn man mit 50 Stundenkilometern durch die Stadt zockelt, spürt man die Kraft, die abgerufen werden könnte.

ES: Das kleine Kind zockelt an der Hand des Vaters hinter ihm her.

Target Word # 166: Seitenaufprall-Schutzsystem (Side Impact Protection System)

OS: Aber wir fahren angstfrei dank der Volvo-Sicherheitskultur ("SIPS Seitenaufprall-Schutzsystem, WHIPS Schleudertrauma-Schutzsystem und ROPS Überschlag-Schutzsystem").

ES: Unser Auto ist mit dem neuesten Seitenaufprall-Schutzsystem ausgestattet.

Target Word # 167: Schleudertrauma-Schutzsystem (Whiplash Impact Pr. System)

OS: Aber wir fahren angstfrei dank der Volvo-Sicherheitskultur ("SIPS Seitenaufprall-Schutzsystem, WHIPS Schleudertrauma-Schutzsystem und ROPS Überschlag-Schutzsystem").

ES: Unser Auto ist mit dem neuesten Schleudertrauma-Schutzsystem ausgestattet.

Target Word # 168: Überschlag-Schutzsystem (Rollover Impact Pr. System)

OS: Aber wir fahren angstfrei dank der Volvo-Sicherheitskultur ("SIPS Seitenaufprall-Schutzsystem, WHIPS Schleudertrauma-Schutzsystem und ROPS Überschlag-Schutzsystem").

ES: Unser Auto ist mit dem neuesten Überschlag-Schutzsystem ausgestattet.

Target Word # 169: bolzen (to hightail, hurtle)

OS: Nein, zum Bolzen ist dieser Wagen nicht geeignet.

ES: Die jungen Männer in ihren schnellen Autos bolzen über die Autobahn.

Target Word # 170: Anmut (grace, charm)

OS: Es gibt diese Art von Kindern, die etwas ausstrahlen, das wie magisch ist, eine ganz besondere Anmut.

ES: Eine Ballettänzerin bewegt sich mit Anmut und Grazie.

Target Word # 171: Hofstaat (royal household)

OS: Auch in Magda Schneiders Haus Mariengrund dreht sich oft eine ganze Welt um "das hübscheste Kind von Berchtesgaden...", das manchmal mit einem einzigen Lächeln seinen gesamten "Hofstaat" regiert.

ES: Ein König regiert seinen Hofstaat.

Target Word # 172: Rummel (fuss, hustle and bustle)

OS: Die "kleine Königin" beim Sonnenbaden, eine etwa sechsjährige Romy, die den Rummel um sie huldvoll genießt, als alle restlos bemüht sind, den Liegestuhl, in dem sie liegt, zu reparieren.

ES: Samstags Nachmittags ist in der Stadt viel Rummel: Es ist dort sehr hektisch; viele Leute laufen hin und her.

Target Word # 173: huldvoll (gracious)

OS: Die "kleine Königin" beim Sonnenbaden, eine etwa sechsjährige Romy, die den Rummel um sie huldvoll genießt, als alle restlos bemüht sind, den Liegestuhl, in dem sie liegt, zu reparieren.

ES: Eine Königin sitzt huldvoll auf dem Thron und schwingt ihr Zepter.

Target Word # 174: innig (dearly, heartfelt)

OS: Romy, Tochter des berühmten Schauspielerehepaares Magda Schneider und Wolf Albach-Retty, hatte trotz der Kriegsjahre und der Scheidung ihrer Eltern eine idyllische Kindheit - und ein inniges Verhältnis zur Mutter, “meine so fabelhafte seelische und moralische Polizei.”

ES: Sie liebt sie heiss und innig.

Target Word # 175: Dirne (prostitute)

OS: “Frankreich befiehlt dir, gesund zu werden”, telegrafierte Jean Cocteau an ihr Krankenbett, kurz vor der bejubelten Pariser Premiere von “Schade, dass sie eine Dirne ist”.

ES: Im Bordell leben Dirnen.

Target Word # 176: schwindelerregend (dizzy, vertiginous)

OS: “Sie war schwindelerregend romantisch”, so ihre Berliner Freundin Christiane Höllger.

ES: Er stieg auf den Turm - hinauf in schwindelerregende Höhen.

Target Word # 177: untrüglich (unmistakable)

OS: “Sissi” wurde von Ernst Marischka mit einem ”feinen Näschen” im gerade wirtschaftlich wiedererwachenden, aufblühenden “Nachkriegs-Deutsch-Österreichland” in die Welt gesetzt - ein epochaler Traum, handwerklich perfekt, mit untrüglichem Instinkt inszeniert, eine unnachahmliche Mischung aus Charme und Sentiment, Musik und Humor.

ES: Frauen haben ein untrügliches Gespür dafür, wenn ihre Männer lügen: sie merken es sofort.

Target Word # 178: behüten (to protect, look after)

OS: Später schrieb die behütetste Jungfrau der Nation an eine Freundin: “Kannst du dir vorstellen, wie das ist, wenn ein ganzes Land auf deine Entjungferung wartet?”

ES: Sie ist ein sehr behütetes Kind: ihre Eltern lassen sie nie allein irgendwo hingehen.

Target Word # 179: Entjungferung (deflowering)

OS: Später schrieb die behütetste Jungfrau der Nation an eine Freundin: “Kannst du dir vorstellen, wie das ist, wenn ein ganzes Land auf deine Entjungferung wartet?”

ES: Früher hielt der Ehemann nach der Hochzeitsnacht, und somit nach der Entjungferung, das blutige Bettlaken aus dem Fenster.

Target Word # 180: lüstern (lewd)

OS: Und so begann sie, die Romanze, von der Öffentlichkeit lüstern und von den Eltern besorgt verfolgt.

ES: Er ist ein lüsterner alter Mann, der junge Frauen in Bars anspricht.

Target Word # 181: unbändig (unruly)

OS: Romy ist unbändig stolz auf ihn.

ES: Der Held hatte unbändige Kraft und konnte so den Bären töten.

Target Word # 182: schlendern (to stroll)

OS: Aber der Mann, der an einem Dezemberabend vor zweieinhalb Jahren die Soi Binthabat Straße in Hua Hin entlang schlenderte, sich neugierig umschaute und schließlich das Lokal ”Checkpoint Charlie” betrat, war dann doch etwas mehr als nur ein Gast.

ES: Die Leute schlendern durch die Strassen und sehen sich die Schaufenster an.

Target Word # 183: meiden (to avoid)

OS: An dieser Straße gibt es jede Menge Bars und Restaurants, die von den Einheimischen gemieden werden.

ES: Ich meide dieses Restaurant: Da ist es schmutzig und das Essen ist schlecht, deshalb gehe ich da nicht hin.

Target Word # 184: Theke (bar)

OS: “Also, das war so”, sagt er, nimmt noch einen Schluck Bier und sucht sich die gemütlichste Position an der Theke.

ES: Peter steht an der Theke und bezahlt noch ein Bier.

Target Word # 185: mies (poor, wretched)

OS: In der Heimat hatte er Tag für Tag auf Märkten Unterwäsche verkauft, mies verdient und sechs Monate im Jahr in der eisigen Luft der deutschen Hauptstadt gefroren.

ES: Peter hat nur wenig Geld, er ärgert sich über sein mieses Gehalt.

Target Word # 186: vermeintlich (alleged)

OS: Nun war er im vermeintlichen Paradies, wo es immer warm ist, wo die Menschen freundlich und die Frauen schön sind.

ES: Uwe ist im vermeintlichen Paradies; wir wissen ja, dass er eigentlich enttäuscht ist.

Target Word # 187: ergehen (to fare)

OS: Ihm sei es so ergangen wie vielen Deutschen in Hua Hin.

ES: Wie ist es dir in der letzten Zeit ergangen, was hast du so gemacht?

Target Word # 188: Semmel (bun)

OS: Hier gibt es mehrere deutsche Restaurants, ein deutsches Reisebüro und einen deutschen Bäcker, der in der Früh frische Semmeln verkauft.

ES: Wieviel Semmeln, ich meine wieviel Brötchen, möchten sie?

Target Word # 189: Kassler (smoked pork chop)

OS: Currywurst kann man in Hua Hin essen, Kassler mit Sauerkraut ist auch kein Problem, und Papa Joe, ein Berliner, macht die besten Steaks.

ES: Ich habe beim Fleischer 1 kg Kassler gekauft, mein Mann isst so gern salziges Schweinefleisch.

Target Word # 190: Fischkutter (small fishing boat)

OS: Im Hafen sieht man die bunten Fischkutter morgens und abends einlaufen und kann die Fischer beim Leeren ihrer Netze beobachten.

ES: Im Hafen liegt ein alter Fischkutter.

Target Word # 191: schäbig (dingy, run-down)

OS: Schäbige Ramschbuden stehen in direkter Nachbarschaft zu schicken Markenläden, die für europäische Verhältnisse immer noch günstige Ware anbieten.

ES: Das ist ein schäbiger Laden: hier ist es schmutzig, und es gibt nur schlechte Qualität.

Target Word # 192: Ramschbude (junk shop)

OS: Schäbige Ramschbuden stehen in direkter Nachbarschaft zu schicken Markenläden, die für europäische Verhältnisse immer noch günstige Ware anbieten.

ES: Dieser Laden ist eine Ramschbude; hier gibt es nur schlechte Qualität.

Target Word # 193: Entsagung (asceticism)

OS: Schon in der Schule lernen die Kinder Bescheidenheit, Entsagung und Demut, und ein mehrwöchiger Klosteraufenthalt im Leben der Männer ist selbstverständlich.

ES: Der Buddhismus lehrt Entsagung: man soll nicht genusssüchtig sein.

Target Word # 194: Demut (humbleness)

OS: Schon in der Schule lernen die Kinder Bescheidenheit, Entsagung und Demut, und ein mehrwöchiger Klosteraufenthalt im Leben der Männer ist selbstverständlich.

ES: Der Buddhismus lehrt auch die Demut: man soll nicht arrogant und egozentrisch sein.

Target Word # 195: Garküche (cookshop)

OS: Also sitzen die Einheimischen im Freien, treffen sich auf den Märkten, am Hafen oder an den Garküchen.

ES: Die Touristen gehen zu einer Garküche und essen dort Reis und Fisch.

Target Word # 196: arg (very)

OS: Wer mit "nein" antwortet, bekommt sofort etwas auf den Teller: Meistens Reis mit Meeresfrüchten und Gemüse, scharf gewürzt, so dass der durchschnittliche Europäer arg ins Schwitzen kommt.

ES: Ihm ist arg heiß, er ist ganz rot im Gesicht und schwitzt sehr.

Target Word # 197: entpuppen (to turn out to be sth./so.)

OS: Doch das Paradies, in dem er der König war, entpuppte sich als Illusion.

ES: Du hast dich als Feigling entpuppt - das habe ich nicht von dir erwartet.

Target Word # 198: Kohle (dough (money))

OS: Die thailändischen Behörden wollen immer nur Kohle, Kohle, Kohle.

ES: Das Auto ist sehr teuer, es hat richtig Kohle gekostet.

Target Word # 199: schröpfen (to fleece)

OS: Die schröpfen die Ausländer.

ES: Die Behörden schröpfen die Ausländer, sie pressen immer mehr Geld von ihnen.

Target Word # 200: knapp (almost)

OS: Von dort in knappen drei Stunden mit Bus oder Expresszug nach Hua Hin.

ES: Dazu brauchen Sie nur eine knappe Stunde, also vielleicht 50 Minuten.

Target Word # 201: Geschöpf (creature)

OS: Sonderbar das Gemisch dieser Menschen: grossbürgerliche und aristokratische Geschöpfe, Studenten mit weissem Stürmer und roter Schnur.

ES: Der christliche Glaube besagt, dass wir alle die Geschöpfe Gottes sind.

Target Word # 202: Stürmer ([type of hat])

OS: Sonderbar das Gemisch dieser Menschen: großbürgerliche und aristokratische Geschöpfe, Studenten mit weissem Stürmer und roter Schnur.

ES: Früher haben Studenten Stürmer auf dem Kopf getragen, heute tragen sie “baseball caps”.

Target Word # 203: Provenienz (origin, provenance)

OS: Die Eleganz der “aristokratischen Geschöpfe” ist der Gepflegtheit der Geschäftsleute frühkapitalistischen Formats, die gesunde Bäuerlichkeit einem dörflich-proletarischen Typus kommunistischer Provenienz gewichen.

ES: In der Hitlerzeit war arische Provenienz von größter Wichtigkeit.

Target Word # 204: hüten (to tend, guard)

OS: Denn obwohl es nun Warschau war, das sich als Hauptstadt bezeichnen durfte, blieb Krakau jene Stadt, in der nationale Symbole gehütet und patriotische Gesten zelebriert, wo Könige gekrönt und nationale Größen bestattet wurden.

ES: Während die Schafe im schottischen Hochland relativ frei leben, werden große Schafherden in Deutschland von einem Schafhirten gehütet.

Target Word # 205: bestatten (to bury)

OS: Denn obwohl es nun Warschau war, das sich als Hauptstadt bezeichnen durfte, blieb Krakau jene Stadt, in der nationale Symbole gehütet und patriotische Gesten zelebriert, wo Könige gekrönt und nationale Größen bestattet wurden.

ES: Tote werden entweder im Krematorium verbrannt oder auf dem Friedhof bestattet.

Target Word # 206: überlegen (superior)

OS: Auch in kommunistischen Zeiten hatte Krakau allen Grund, sich Warschau gegenüber überlegen zu fühlen:

ES: Der Sieger des Rennens hatte sich viel besser vorbereitet als seine Gegner und war ihnen deshalb klar überlegen.

Target Word # 207: Glaubensbekenntnis (creed)

OS: “Der Umstand, in einer der beiden Städte zu wohnen, kommt fast schon einem Glaubensbekenntnis gleich”, schrieb in den siebziger Jahren der deutsche Schriftsteller Rolf Schneider.

ES: Wenn junge Menschen das Trikot von Celtic oder Rangers tragen, kommt das einem Glaubensbekenntnis gleich.

Target Word # 208: Ruch (reputation)

OS: Die Wahl für Krakau ist umgeben vom Ruch des Snobismus, vermengt mit zarter Provinzialität.

ES: Oft sieht man einem Millionär an, dass er viel Geld hat, weil er vom Ruch des Geldes umgeben ist.

Target Word # 209: rau (cragged, gnarly)

OS: Die Wahl für Warschau wird als Zeichen für raue Manieren, schneidenden Ehrgeiz und Wurstigkeit gegenüber feiner alter Kultur angesehen.

ES: Wenn man oft ohne Schutzhandschuhe im Garten arbeitet und seine Hände nicht cremt, bekommt man raue Hände.

Target Word # 210: Wurstigkeit (indifference)

OS: Die Wahl für Warschau wird als Zeichen für raue Manieren, schneidenden Ehrgeiz und Wurstigkeit gegenueber feiner alter Kultur angesehen.

ES: Er erledigte den Job mit einer Wurstigkeit, die jedem zeigte, wie egal und unwichtig ihm dieser Job war.

Target Word # 211: draufgängerisch (ballsy)

OS: Es wird ihnen immer noch nachgesagt, sie seien aufbrausend, draufgängerisch, mitunter aggressiv.

ES: Man sagt, dass die Einwohner Warschaus draufgängerisch sind, weil sie oft viel riskieren.

Target Word # 212: verlässlich (reliable)

OS: Angeblich verbirgt sich dahinter die Sehnsucht nach Tradition und verlässlichen Werten.

ES: Eine Fahrt mit dem Auto durch die Wüste Sahara ist ein großes Abenteuer, aber dazu braucht man ein verlässliches Auto, z.B. einen Landrover.

Target Word # 213: spotten (to mock)

OS: Der Warschauer spotte gern über den konservativen Traditionalismus Krakaus, behauptete einmal der Publizist Krzysztof T. Toeplitz, aber er beneide es insgeheim darum, dass “Sessel oder Stuhl in Krakauer Wohnungen Möbelstücke vom Großvater oder Urgroßvater sind und das Cafe am Marktplatz wirklich jenes Cafe ist, wo die künstlerische Boheme des Fin de Siècle zu sitzen pflegte”.

ES: Egal was er macht, alle spotten nur über ihn, weil er nichts richtig machen kann.

Target Word # 214: beneiden (to envy)

OS: Der Warschauer spotte gern über den konservativen Traditionalismus Krakaus, behauptete einmal der Publizist Krzysztof T. Toeplitz, aber er beneide es insgeheim darum, dass “Sessel oder Stuhl in Krakauer Wohnungen Möbelstücke vom Großvater oder Urgroßvater sind und das Cafe am Marktplatz wirklich jenes Cafe ist, wo die künstlerische Boheme des Fin de Siècle zu sitzen pflegte”.

ES: Obwohl mein Freund viel mehr Geld hat als ich, beneide ich ihn nicht, weil Geld allein nicht glücklich macht.

Target Word # 215: verlegen (to relocate, move)

OS: Viele Krakauer, die in der Hauptstadt beschäftigt sind, denken nicht daran, auch ihren Wohnsitz dorthin zu verlegen.

ES: Wenn ich eine neue Arbeitsstelle in Edinburg bekomme, werde ich meinen Wohnsitz von Dundee nach Edinburg verlegen.

Target Word # 216: säumen (to seam)

OS: Zwar ist auch der Warschauer “Königsweg” nicht nur von Palästen, Kirchen und Denkmälern, sondern auch von Restaurants und Cafes gesäumt, doch die Gelassenheit, die der Krakauer Hauptmarkt verströmt, will sich hier nicht so recht einstellen.

ES: Als Königin Elisabeth anlässlich ihres goldenen Jubiläums durch London fuhr, waren die Strassen von Touristen gesäumt.

Target Word # 217: mäßig (moderate)

OS: Allein die Krakauer Vorstadt enthält, wie einst J. C. F. Schulz in seiner “Reise nach Warschau” notierte, “in einer mäßigen Länge elf Paläste, worunter einige sind, deren sich der mächtigste regierende Fürst nicht schämen dürfte”.

ES: Mit ihrer letzten CD hatte die Gruppe “Oasis” nur mäßigen Erfolg im Vergleich zu ihren Erfolgen vor einigen Jahren.

Target Word # 218: schämen (to be ashamed)

OS: Allein die Krakauer Vorstadt enthält, wie einst J. C. F. Schulz in seiner “Reise nach Warschau” notierte, ”in einer mäßigen Länge elf Paläste, worunter einige sind, deren sich der mächtigste regierende Fürst nicht schämen dürfte”.

ES: Wenn man einen Fehler macht, sollte man sich nicht schämen, aber man sollte versuchen, es das nächste Mal besser zu machen.

Target Word # 219: Schwermut (gloom)

OS: Auf der einen Seite das “Europejski”, dem immer noch ein Hauch des Fin de Siècle und zugleich eine Mischung aus slawischer Schwermut und postkommunistischer Tristesse anhaften.

ES: Seine Stimmung schwankte immer zwischen den 2 Extremen Euphorie und Schwermut.

Target Word # 220: ausgerechnet (just, of all things/persons)

OS: Ausgerechnet dieser Stolz der kommunistischen Machthaber, für dessen Bau die Parteimitglieder freiwillig Geld spendeten, wurde Anfang der neunziger Jahre zu einem “Bank- und Finanzzentrum” umfunktioniert, dessen Hauptteil die neu entstandene Warschauer Börse bildete.

ES: Er war auf vieles vorbereitet, aber dass er ausgerechnet seine Exfrau in der Dating-Agentur treffen würde, damit hatte er nicht gerechnet.

Target Word # 221: spenden (to donate)

OS: Ausgerechnet dieser Stolz der kommunistischen Machthaber, für dessen Bau die Parteimitglieder freiwillig Geld spendeten, wurde Anfang der neunziger Jahre zu einem “Bank- und Finanzzentrum” umfunktioniert, dessen Hauptteil die neu entstandene Warschauer Börse bildete.

ES: Das Rote Kreuz konnte vielen hungernden Menschen in Afrika helfen, weil viele Leute sehr viel Geld spendeten.

Target Word # 222: münden (to flow, lead into)

OS: Jenseits des Platzes der Drei Kreuze, in den die Neue Welt mündet, beginnt die Ujazdowski-Allee, eine Diplomatenmeile mit zahlreichen Botschaften und Regierungsgebäuden.

ES: Der Fluss “Tay” mündet in die Nordsee.

Target Word # 223: Entwurzelung (uprooting, rootlessness)

OS: Die turbulente Handlung täuscht kaum über das eigentliche Thema hinweg: die Tragik dieser Stadt und die Entwurzelung ihrer Bewohner, die seit Jahrzehnten “die imaginäre Eroberung der Innenstadt mit ihren Wundern, ihrem Glanz und ihrer Pracht proben”.

ES: Das Problem von Bürgerkriegsflüchtlingen ist nicht nur, dass sie fast alles verlieren, was sie besitzen, sondern vor allem der Verlust ihrer Heimat, ihre völlige Entwurzelung.

Target Word # 224: nicht mal (not even)

OS: Da gab es keine Rohstoffe, keine Reichtümer, nicht mal Wohlstand, nichts, was sich zu erobern lohnte.

ES: Ich war vollkommen pleite, und konnte nicht mal meine Frau anrufen.

Target Word # 225: Herrensitz (manor)

OS: Einen Herrensitz mit Siedlung nannten sie Spandow.

ES: König Ludwig ließ noch einen prächtigen Herrensitz am Chiemsee bauen.

Target Word # 226: Geschlecht (clan, dynasty)

OS: Dieser Askanierfürst ermunterte Angehörige seines Geschlechts, aber auch Landsuchende aus Franken und dem Rheinland, hierher zu kommen.

ES: In Schottland trägt man oft einen Schottenrock mit dem traditionellen Muster des Geschlechts, dem man angehört.

Target Word # 227: bekehren (to convert)

OS: Die Landnahme erfolgt friedlich, die Bevölkerung wurde zum Christentum bekehrt.

ES: Mein Bruder ist Missionar in Afrika und ist stolz, mehrere Einheimische zum Christentum bekehrt zu haben.

Target Word # 228: urkundlich (documentary)

OS: Im Jahre 1237 jedenfalls wird die städtische Siedlung Cölln auf der Spreeinsel erstmals urkundlich erwähnt.

ES: Man kann nicht genau nachweisen, wann Berlin gegründet wurde, da die Siedlung dort erst 1237 urkundlich erwähnt wurde.

Target Word # 229: aktenkundig (on record)

OS: Obwohl Berlin erst sieben Jahre später aktenkundig ist.

ES: Die Polizei nahm die Information entgegen, und kurz darauf wurde der Fall endlich aktenkundig.

Target Word # 230: gedeihen (to thrive, prosper)

OS: Das Gemeinwesen in den “Schwesterstädten” gedieh, sie besassen einen gemeinsamen Rat, erwarben 1369 vom Landesherrn das eigene Münzrecht und waren ansonsten stolz auf ihre Eigenständigkeit.

ES: Ich bin froh, dass meine Tochter wächst und gedeiht.

Target Word # 231: Zwingburg (fortress, stronghold)

OS: Er ließ auf der Insel eine Zwingburg errichten.

ES: Im Mittelalter war es üblich, dass ein Herrscher eine Zwingburg bauen ließ, um die Bürger einzuschüchtern.

Target Word # 232: Söldner (mercenary)

OS: 1713 beginnt Friedrich Wilhelm I., genannt der Soldatenkönig, mit dem Aufbau einer grossen eigenen Armee aus Söldnern.

ES: Das Land hatte keine eigene Armee und musste Söldner aus verschiedenen Ländern anheuern.

Target Word # 233: Zeughaus (armory)

OS: Ein barockes Zeughaus, das Armen-Hospital Charite, Opernhaus, Schauspielhaus, Porzellanmanufaktur, später eine Universität.

ES: Der König ließ ein Zeughaus bauen, wo er neben dem Palast seine Wache unterbrachte.

Target Word # 234: Ausrottung (eradication)

OS: In Deutschland nahm der Antisemitismus zu, und die Ausrottung aller Juden wurde gefordert.

ES: Weil die Ausrottung der Juden den Nationalsozialisten nicht gelungen ist, gibt es immer noch Juden in Zentraleuropa.

Target Word # 235: dulden (to tolerate)

OS: Und: Kaiser und Kirche duldeten das.

ES: Ich bin zwar tolerant, aber Rassismus werde ich in meiner Gegend nicht dulden.

Target Word # 236: Größenwahn (megalomania)

OS: Der Weltkrieg 1914-18 endete in einem Desaster des Größenwahns.

ES: Hitler wollte die ganze Welt erobern und litt offensichtlich unter Größenwahn.

Target Word # 237: partout (absolutely)

OS: Die Welt wollte partout nicht am deutschen Wesen genesen.

ES: Da ich die Sonne hasse, wollte ich partout nicht in den Süden fliegen.

Target Word # 238: Pflaster (pavement, place [fig.])

OS: Ein gutes Pflaster für die Nationalsozialisten unter einem, der sich Führer nannte und Adolf Hitler hieß, zumal sie auf das fußen konnten, was vorher hier erdacht worden war, Herrenrasse und Judenvernichtung.

ES: Russland im Jahre 1918 war ein gutes Pflaster für die Kommunisten, da das System gegen die Bürger gerichtet war.

Target Word # 239: schändlich (ignoble, disgraceful)

OS: Es begann die blutigste und schändlichste Zeit deutscher Geschichte mit 50 Millionen Toten in Europa, in Konzentrationslagern und Gaskammern.

ES: Der Westen hat der Welt viel Gutes gebracht, aber es gibt auch schändliche Episoden in seiner Geschichte.

Target Word # 240: uneingeschränkt (unlimited)

OS: Politiker aus allen Regionen der westlichen Welt zeigten Betroffenheit, John F. Kennedy erklärte im Juni 1963 seine uneingeschränkte Solidarität mit den (deutsch gesprochenen) Worten: “Ich bin ein Berliner”.

ES: Wenn man eine Broadband-Internet-Verbindung hat, hat man uneingeschränkten Zugang zum Internet.

Target Word # 241: brodeln (to seethe)

OS: Trotzdem brodelte es im ökonomisch kränkelnden Ostberlin, führten wirtschaftliche und politische Unzufriedenheit in der DDR zu wachsendem Volkszorn.

ES: Die Arbeiter waren unzufrieden, und man sah förmlich, wie es vor dem Streik in der Fabrik brodelte.

Target Word # 242: Volkszorn (public outrage)

OS: Trotzdem brodelte es im ökonomisch kränkelnden Ostberlin, führten wirtschaftliche und politische Unzufriedenheit in der DDR zu wachsendem Volkszorn.

ES: Der Volkszorn in Russland führte 1918 zur Revolution.

Target Word # 243: verzichten (to do without, forgo)

OS: In den Zwei-plus-Vier-Verhandlungen vom 12.September 1990 verzichteten die einstigen Siegermächte auf ihren Sonderstatus in Berlin.

ES: Ich habe das Recht von Ihnen eine Entschuldigung zu bekommen, aber ich verzichte darauf.

Appendix C

Syntactic Predictions for SPLT

- *Matrix verbs* are cost-free except in the case of an additional main clause joined by coordination and are predicted at the coordinating conjunction at the earliest;
- *Matrix subjects* are always predicted at the beginning of a sentence except in the case of an additional main clause joined by coordination, which is predicted at the coordinating conjunction at the earliest;
- *Null subjects* (matrix or relative clause) (e.g. *Ihm ist kalt* [he is cold]) are treated as if they were predicted (but never realized) from the start until the end of the sentence or relative clause;
- *Matrix objects* are predicted at the corresponding transitive verb or later; in general, indirect (dative) objects are not predicted as they are usually optional and not introduced by a preposition;
- *Matrix subject complements* (predicate adjectives or noun complements, PP or passive complements, complement-clauses) are predicted at a finite form of *sein, wirken* etc [to be, seem]) or later in the case of an additional complement;
- *Infinite Verb(group)s* are predicted at finite auxiliary/modal verbs;
- *Relative Clause/Apposition Subjects* are predicted at the beginning of a relative clause/apposition;
- *Relative Clause/Apposition Verbs* are predicted at the start of a relative clause/ apposition; a coordination particle after verb (*und, oder* [and, or]) predicts an additional RC verb;

- *"zu"-infinitives* are predicted at finite modality verbs such as *versuchen* [to try], *anfangen* [to begin] etc.;
- *Prepositional object in relative clauses* are predicted at introducing prepositions;
- *Complements of verbal infinitives* (e.g. *mitanzuhören, wie er schrie* [to listen how he cried]) are predicted at verbal infinitives;
- *Separated Verb Particles (Prefix or Complement)* are predicted at the corresponding verb stems;
- *Clause-like participle constructions* if fronted by participle are predicted at that participle (e.g. *Kombiniert mit Dr. Stoibers Schweigen* [combined with Dr. Stoiber's silence]).

Appendix D

Selected Sentence Pairs for Sentence Similarity Study

The valid target word are given in italics; sentence pairs *not* part of the final set of teacher data due to their inclusion of multi-word target words are marked with an asterisk (*).

Sentence Pair # 1:
OS: Fleißige Klosterschüler bekamen die Backware als *Ansporn* für das Lernen neuer Gebete.
ES: Kindern gibt man kleine Belohnungen als *Ansporn* zum Lernen.

Sentence Pair # 2*:
OS: Seine Landsleute nahmen ihm das Missgeschick gern ab.
ES: Diese Geschichte nehme ich dir gerne ab.

Sentence Pair # 3:
OS: Die Baltimore Ravens, die Titelverteidiger mit ihrer bärenstarken Abwehr, beherrschten die Miami Dolphins *nach Belieben* und gewannen vernichtend mit 20:3.
ES: Die Musiker spielten *nach Belieben*, bis der Dirigent auf das Podium trat.

Sentence Pair # 4:

OS: Jeder, der bei *Kampfhandlungen* festgenommen wird, gilt als PoW und hat sofortigen und vollständigen Anspruch auf Schutz durch die Genfer Konvention (zumindest bis ein "kompetentes Tribunal" seinen endgültigen Status klärt).

ES: Jeder, der bei *Kampfhandlungen* in Afghanistan festgenommen wird, gilt als PoW.

Sentence Pair # 5:

OS: Die meisten Politiker und Juristen *fressen* ihren Groll still in sich hinein.

ES: Hunde *fressen* gerne rohes Fleisch.

Sentence Pair # 6:

OS: Noch immer *verübelt* Gennifer Flowers Clinton, dass er nicht offen zu der Affäre stand, sondern sie 'verraten' hat.

ES: Er hat mir sehr *verübelt*, dass ich bei der Auseinandersetzung nicht für ihn Partei ergriffen habe.

Sentence Pair # 7:

OS: Zehn Jahre später traf Stern-Reporter Claus Lutterbeck eine *rachsüchtige* Ex-Geliebte in New Orleans.

ES: Sie war sehr *rachsüchtig* und versuchte bei jeder Gelegenheit, ihm das Böse, das er ihr angetan hatte, heimzuzahlen.

Sentence Pair # 8:

OS: Wie hemmungslos Mitarbeiter im Öffentlichen Dienst ihre Rechner für ihr Freizeitvergnügen nutzen, belegt jetzt eine Untersuchung des Niedersächsischen *Landesrechnungshofs.*

ES: Der *Landesrechnungshof* hat geprüft, ob die Beamten im Ministerium ihre Aufgaben richtig erfüllen.

Sentence Pair # 9:

OS: Für ihre Studie werteten die Kontrolleure erstmals die *Zugriffe* von 20 000 Landesbediensteten aus, die sich über das Informatikzentrum Niedersachsen ins Internet einwählen.

ES: Je mehr *Zugriffe* meine Webseite aufweist, desto sicherer bin ich, dass die Leute sich für sie interessieren.

Sentence Pair # 10:

OS: Andreas B., 35, *ledig*, kinderlos, war zu Recht fristlos gefeuert worden, befanden die Richter.

ES: Ein Bruder von mir ist verheiratet, aber der andere ist noch *ledig*.

Sentence Pair # 11:

OS: Sollte El Niño die Wetterküche auch in diesem Jahr aufwühlen, würden erneut vor allem die Armen der Welt die *Zeche* bezahlen.

ES: Sie verließen die Bar ohne die *Zeche* zu bezahlen.

Sentence Pair # 12:

OS: 22 000 Menschen kamen in Sturmfluten und Feuersbrünsten um, sie verhungerten, weil die Ernte auf den Feldern verdarb, oder sie fielen *Seuchen* zum Opfer, die sich im Gefolge des Wetterdurcheinanders ausbreiteten.

ES: Im Mittelalter war die Pest eine verbreitete *Seuche*.

Sentence Pair # 13:

OS: Rätselhaft ist, warum die seit mindestens 130 000 Jahren *nachweisbare* Unwetter-Konstellation neuerdings immer häufiger und stärker über den Planeten hereinbricht.

ES: Er hatte Glück, denn im Blut war kein Alkohol *nachweisbar*.

Sentence Pair # 14:

OS: Bei manchen kommt das Gefühl der Angst immer wieder - regelmäßig und *zerstörerisch*.

ES: Fußballfans sind oft *zerstörerisch*, nachdem sie verloren haben: sie machen dann Dinge kaputt.

Sentence Pair # 15*:

OS: Nach dem 11.September, so ergab eine Studie des *Berufsverbands* der Allgemeinärzte, erschienen rund 45 Prozent mehr Patienten mit Angststörungen in den deutschen Praxen.

ES: Ein *Berufsverband* ist ein Zusammenschluß oder eine Gruppe von Menschen, die denselben Beruf haben.

Sentence Pair # 16:

OS: Aber die Niederlage ist dramatisch und *bedrückend*.

ES: Die Hitze ist *bedrückend*.

Sentence Pair # 17:

OS: Ich glaube, es fängt schon mit einem bestimmten Grundton an, dem Ton der *Häme*.

ES: Er war schadenfroh wie immer und betrachtete ihr gebrochenes Bein voller *Häme*.

Sentence Pair # 18:

OS: Und dann kann die ganze Kollisionswarntechnik nicht funktionieren, sind die Piloten allein auf die Bodenkontrolle *angewiesen*.

ES: Die Piloten müssen der Bodenkontrolle vertrauen, weil sie ganz auf sie *angewiesen* sind.

Sentence Pair # 19:

OS: Tornados fliegen über den Absturzkorridor und machen Luftaufnahmen, damit alle Opfer so schnell wie möglich gefunden und *geborgen* werden können.

ES: Nach dem Schiffsunglück versuchen nun Rettungsteams, die Opfer zu *bergen*.

Sentence Pair # 20*:

OS: Diesem Befehl, das entspricht der international geltenden Absprache, muss der Pilot unverzüglich und ohne weitere Absprache *Folge leisten.*

ES: Du musst mir unbedingt *Folge leisten* und tun, was ich dir sage, damit nichts schief geht.

Sentence Pair # 21:

OS: Mit ihren verseuchten Löwen und Tigern, mit mehr als 177 000 Rindern und Kühen, denen BSE in den vergangenen zwölf Jahren zum Verhängnis wurde, und mit inzwischen über 80 Menschen, die vCJK, die menschliche Variante des Rinderwahns, das Leben kostete, haben die Bewohner des Vereinigten Königreichs der BSE-Epidemie einen hohen Preis *gezollt*.

ES: Angesichts der vielen Todesfälle bei Mensch und Tier haben die Briten der Epidemie einen hohen Preis *gezollt*.

Sentence Pair # 22:

OS: Auch an vielen britischen Schulen, die in den 90er Jahren Rindfleisch aus ihren Küchen *verbannten*, taucht es mittlerweile wieder auf der Speisekarte auf.

ES: Als Napoleon *verbannt* wurde, ging er nach Elba ins Exil.

Sentence Pair # 23*:

OS: Die Kultur des Ressentiments appelliert an Gefühle und Ängste, ohne sich eine Blöße zu geben oder sich ertappen zu lassen.

ES: Der Dieb war so dumm und machte so viel Lärm, dass er sich von der Polizei ertappen ließ.

Sentence Pair # 24:

OS: Und wie kann man da populär werden, ohne einerseits als Entertainer zu glänzen und andererseits mit *Augenblinzeln* und Andeutungen - also ohne das Vulgäre von Westerwelle & Möllemann - auch jenen ein Obdach zu bieten, die sich in den großen Parteien derzeit nicht mehr zu Hause fühlen?

ES: Die Frau kommunizierte mit dem attraktiven Mann in der Bar mit *Augenblinzeln*.

Sentence Pair # 25*:

OS: Gleichzeitig brachte Blair die Gewerkschaften daheim nachhaltig gegen sich auf, indem er mit Silvio Berlusconi eine Allianz gegen Frankreich und Deutschland begründete, um die Arbeitnehmerrechte in der EU zu schwächen.

ES: Alle waren ärgerlich auf ihn, denn er hatte alle durch sein unsoziales Verhalten gegen sich aufgebracht.

Sentence Pair # 26:

OS: Seit sich der Premier dem Krieg gegen den Terrorismus verschrieben hat, war die Heimatfront noch nie so *wacklig*.

ES: Der Betrunkene steht auf *wackligen* Beinen.

Sentence Pair # 27:

OS: Gleichzeitig brachte Blair die Gewerkschaften daheim *nachhaltig* gegen sich auf, indem er mit Silvio Berlusconi eine Allianz gegen Frankreich und Deutschland begründete, um die Arbeitnehmerrechte in der EU zu schwächen.

ES: Die Maßnahmen sollten nicht nur kurzfristig wirken, sondern *nachhaltig* sein.

Sentence Pair # 28:

OS: Tage später steckte meine Frau Doris in einem künstlich erzeugten Autostau, in dem in letzter Sekunde eine präparierte Bombe *entschärft* wurde.

ES: Die Bombe wurde rechtzeitig *entschärft* - fünf Minuten später wäre sie explodiert.

Sentence Pair # 29:

OS: Die besonders scharf bewachte amerikanische Schule in Herzlia ist wohl der sensibelste *Gradmesser* für die politische Situation.

ES: Die guten Schulnoten beweisen die hohe Qualität des Unterrichts: Sie sind ein Gradmesser, also ein klarer Indikator dafür.

Sentence Pair # 30*:

OS: Und die nervenden Geräusche kommen jetzt nicht vom Dach, sondern von den anderen Rasern, die neben uns herheulen.

ES: Das andere Auto heult so laut neben mir her, dass ich nicht mehr Radio hören kann.

Sentence Pair # 31*:

OS: Selbst wenn man mit 50 Stundenkilometern durch die Stadt zockelt, spürt man die Kraft, die abgerufen werden könnte.

ES: Die benötigten Informationen können auf unserer Website abgerufen werden.

Sentence Pair # 32*:

OS: “Ich wäre meiner Mutter mit vierzehn durchgebrannt, wenn sie mich nicht hätte zum Film gelassen”, sagt Romy später.

ES: Sie war verheiratet, aber sie ist mit ihrem Liebhaber durchgebrannt.

Sentence Pair # 33:

OS: “Frankreich befiehlt dir, gesund zu werden”, telegrafierte Jean Cocteau an Romy Schneiders Krankenbett, kurz vor der bejubelten Pariser Premiere von “Schade, dass sie eine *Dirne* ist”.

ES: Im Bordell leben *Dirnen*.

Sentence Pair # 34*:

OS: Vom Urlaub in den Alltag: In Thailands Badeort Hua Hin lassen sich immer mehr Deutsche nieder.

ES: In Hua Hin wohnen viele Deutsche, die sich da in den letzten paar Jahren niedergelassen haben.

Sentence Pair # 35:

OS: "Also, das war so", sagt Wirt Uwe Dörk, nimmt noch einen Schluck Bier und sucht sich die gemütlichste Position an der *Theke*.

ES: Peter steht an der *Theke* und bezahlt noch ein Bier.

Sentence Pair # 36:

OS: Es wird den Warschauern immer noch nachgesagt, sie seien aufbrausend, *draufgängerisch*, mitunter aggressiv.

ES: Man sagt, dass die Einwohner Warschaus *draufgängerisch* sind, weil sie oft viel riskieren.

Sentence Pair # 37:

OS: Auf der einen Seite ist da das Hotel "Europejski", dem immer noch ein Hauch des Fin de Siècle und zugleich eine Mischung aus slawischer *Schwermut* und postkommunistischer Tristesse anhaften.

ES: Seine Stimmung schwankte immer zwischen den 2 Extremen Euphorie und *Schwermut*.

Sentence Pair # 38:

OS: Die Wahl für Warschau wird als Zeichen für *raue* Manieren, schneidenden Ehrgeiz und Wurstigkeit gegenüber feiner alter Kultur angesehen.

ES: Wenn man oft ohne Schutzhandschuhe im Garten arbeitet und seine Hände nicht cremt, bekommt man *raue* Hände.

Sentence Pair # 39:

OS: Trotzdem brodelte es im ökonomisch kränkelnden Ostberlin, führten wirtschaftliche und politische Unzufriedenheit in der DDR zu wachsendem *Volkszorn*.

ES: Der *Volkszorn* in Russland führte 1918 zur Revolution.

Sentence Pair # 40:

OS: Die Welt wollte *partout* nicht am deutschen Wesen genesen.

ES: Da ich die Sonne hasse, wollte ich *partout* nicht in den Süden fliegen.

Appendix E

Instructions for Sentence Similarity Study

Sentence Similarity Study

Many thanks for doing our survey!

Please read through the instructions below carefully before starting. If you have any questions or observations about the study, please get in contact with us – we'd be delighted to hear from you.

Instructions

You will be given 2 booklets of 20 sentence pairs to read. For each of the sentence pairs, your task is to judge the semantic similarity of the two sentences, on a 10 point scale. On this scale, 1 denotes the lowest and 10 the highest degree of similarity.

Suppose you were asked to rate the following sentence pair:

Der Kanzler entfernte sich gestern schnell.

(The chancellor went away quickly yesterday.)

Gerhard Schröder hatte es gestern eilig wegzukommen.

(Gerhard Schröder was in a hurry to get away yesterday.)

These 2 sentences are quite similar, so you would probably give them a relatively high number.

Now consider the following sentence pair:

Im Mittelalter hatten die Fürsten Vasallen, die ihnen in jeder Situation treu sein mussten und dafür von ihnen beschützt und immer gut bezahlt wurden.

(In medieval ages, princes had vassals who had to be faithful to them and were protected and always well paid by them [the princes] in return.)

Dienstags geht Brigitte oft ins Kino, obwohl sie sich das zeitlich nicht immer leisten kann.

(Brigitte often goes to to the movies on Tuesdays, even though she cannot always afford the time for it.)

These 2 sentences are not very similar, so you would probably give them a relatively low number.

There are no 'correct' answers, so whatever number seems appropriate to you is a valid response. While you are deciding a number for a similarity rating, please bear in mind the following:

- It is only the semantic similarity of the sentence pairs that you should consider, ie. the degree to which they "mean" or "are about" the same thing(s), not their syntactic (structural) similarity;
- All sentence pairs have at least one word in common; please give the sentence pair that you think has the lowest similarity a 1 regardless.

IMPORTANT!!!

Please strictly follow the following steps for this study:

1. For each booklet, please read all 20 sentence pairs first before rating any of them;

2. Next, choose the sentence pair that you think has the *lowest similarity* among the sentence pairs in the booklet, and give it a 1;

3. Next, choose the sentence pair that you think has the *highest* similarity among the sentence pairs, and give it a 10;

4. Now rate the remaining 18 sentence pairs on the 10 point scale (possibly, but not necessarily, including the extreme values 1 and 10).

The survey will take approximately 20-30 minutes. Thanks for taking part! You can start the study proper by pressing on the 'Start' button below. The sentences will appear in a separate window, so you can still look at the instructions while rating, if you wish. Please make sure that javascript is enabled in your browser. When you have finished rating the first booklet of 20 sentence pairs, please click on the 'Continue' button at the bottom of the window to rate the second booklet of sentence pairs. However, you also have the option to exit the study at this point and submit your ratings for the first booklet only - in this case, please click on the 'Submit' button (also at the bottom of the window).

Appendix F

Questionnaire (Main Section) for Sentence Similarity Study

Personal Details

Name:

E-mail:

1

Und die nervenden Geräusche kommen jetzt nicht vom Dach, sondern von den anderen Rasern, die neben uns herheulen.
Das andere Auto heult so laut neben mir her, dass ich nicht mehr Radio hören kann.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

2

Nach dem 11.September, so ergab eine Studie des Berufsverbands der Allgemeinärzte, erschienen rund 45 Prozent mehr Patienten mit Angststörungen in den deutschen Praxen.
Ein Berufsverband ist ein Zusammenschluß oder eine Gruppe von Menschen, die denselben Beruf haben.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

3

Seine Landsleute nahmen ihm das Missgeschick gern ab.
Diese Geschichte nehme ich dir gerne ab.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

4

Wie hemmungslos Mitarbeiter im Öffentlichen Dienst ihre Rechner für ihr Freizeitvergnügen nutzen, belegt jetzt eine Untersuchung des Niedersächsischen Landesrechnungshofs.
Der Landesrechnungshof hat geprüft, ob die Beamten im Ministerium ihre Aufgaben richtig erfüllen.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

5

Jeder, der bei Kampfhandlungen festgenommen wird, gilt als PoW und hat sofortigen und vollständigen Anspruch auf Schutz durch die Genfer Konvention (zumindest bis ein "kompetentes Tribunal" seinen endgültigen Status klärt).
Jeder, der bei Kampfhandlungen in Afghanistan festgenommen wird, gilt als PoW.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

6

Rätselhaft ist, warum die seit mindestens 130 000 Jahren nachweisbare Unwetter-Konstellation neuerdings immer häufiger und stärker über den Planeten hereinbricht.
Er hatte Glück, denn im Blut war kein Alkohol nachweisbar.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

7

Tornados fliegen über den Absturzkorridor und machen Luftaufnahmen, damit alle Opfer so schnell wie möglich gefunden und geborgen werden können.
Nach dem Schiffsunglück versuchen nun Rettungsteams, die Opfer zu bergen.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

8

Selbst wenn man mit 50 Stundenkilometern durch die Stadt zockelt, spürt man die Kraft, die abgerufen werden könnte.
Die benötigten Informationen können auf unserer Website abgerufen werden.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

9

Die Welt wollte partout nicht am deutschen Wesen genesen.
Da ich die Sonne hasse, wollte ich partout nicht in den Süden fliegen.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

10

Diesem Befehl, das entspricht der international geltenden Absprache, muss der Pilot unverzüglich und ohne weitere Absprache Folge leisten.
Du musst mir unbedingt Folge leisten und tun, was ich dir sage, damit nichts schief geht.

lowest similarity	1	2	3	4	5	6	7	8	9	10	highest similarity

11

Da ich nicht mehr im Getümmel stecke, richte ich mich nach den Meinungsumfragen.
Beim Sommerschlussverkauf waren so viele Menschen, dass ich dem Getümmel meine Freundin verlor.

lowest similarity	1	2	3	4	5	6	7	8	9	10	highest similarity

12

Trotzdem brodelte es im ökonomisch kränkelnden Ostberlin, führten wirtschaftliche und politische Unzufriedenheit in der DDR zu wachsendem Volkszorn.
Der Volkszorn in Russland führte 1918 zur Revolution.

lowest similarity	1	2	3	4	5	6	7	8	9	10	highest similarity

13

Sollte El Niño die Wetterküche auch in diesem Jahr aufwühlen, würden erneut vor allem die Armen der Welt die Zeche bezahlen.
Sie verließen die Bar ohne die Zeche zu bezahlen.

lowest similarity	1	2	3	4	5	6	7	8	9	10	highest similarity

14

Tage später steckte meine Frau Doris in einem künstlich erzeugten Autostau, in dem in letzter Sekunde eine präparierte Bombe entschärft wurde.
Die Bombe wurde rechtzeitig entschärft - fünf Minuten später wäre sie explodiert.

lowest similarity	1	2	3	4	5	6	7	8	9	10	highest similarity

15

22 000 Menschen kamen in Sturmfluten und Feuersbrünsten um, sie verhungerten, weil die Ernte auf den Feldern verdarb, oder sie fielen Seuchen zum Opfer, die sich im Gefolge des Wetterdurcheinanders ausbreiteten.
Im Mittelalter war die Pest eine verbreitete Seuche.

lowest similarity	1	2	3	4	5	6	7	8	9	10	highest similarity

16

Die Baltimore Ravens, die Titelverteidiger mit ihrer bärenstarken Abwehr, beherrschten die Miami Dolphins nach Belieben und gewannen vernichtend mit 20:3.
Die Musiker spielten nach Belieben, bis der Dirigent auf das Podium trat.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

17

Die Wahl für Warschau wird als Zeichen für raue Manieren, schneidenden Ehrgeiz und Wurstigkeit gegenüber feiner alter Kultur angesehen.
Wenn man oft ohne Schutzhandschuhe im Garten arbeitet und seine Hände nicht cremt, bekommt man raue Hände.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

18

Die Kultur des Ressentiments appelliert an Gefühle und Ängste, ohne sich eine Blöße zu geben oder sich ertappen zu lassen.
Der Dieb war so dumm und machte so viel Lärm, dass er sich von der Polizei ertappen ließ.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

19

"Also, das war so", sagt Wirt Uwe Dörk, nimmt noch einen Schluck Bier und sucht sich die gemütlichste Position an der Theke.
Peter steht an der Theke und bezahlt noch ein Bier.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

20

Mit ihren verseuchten Löwen und Tigern, mit mehr als 177000 Rindern und Kühen, denen BSE in den vergangenen zwölf Jahren zum Verhängnis wurde, und mit inzwischen über 80 Menschen, die vCJK, die menschliche Variante des Rinderwahns, das Leben kostete, haben die Bewohner des Vereinigten Königreichs der BSE-Epidemie einen hohen Preis gezollt.
Angesichts der vielen Todesfälle bei Mensch und Tier haben die Briten der Epidemie einen hohen Preis gezollt.

lowest similarity	○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10	highest similarity

Submit and Exit | Clear Form | Continue with next 20 Sentences

Appendix G

Test Items for the Multiple-Choice Lexical Relations Test

In the following listings, the target word is given in boldface, followed by the underlined word that is semantically related to the target word, and a set of distractors. Of these, deliberately chosen 'misleading' distractors are marked with an asterisk; the rest are random distractors.

1.1 Nouns - Synonyms

Missgeschick (*mishap*)	**Blessuren** (*wounds*)	**Anschlag** (*strike*)
Ungeschick	Verwundungen	Attentat
Rausschmeißer	Klausuren*	Abschlag*
Wanderstempel	Konturen	Schlag*
Schnellzug	Sozialminister	Katholizismus
Indianerin	Funde	Kuchenteller
Finanzministerium	Frauenbewegungen	Aufwärmtraining
Feuerwehrhaus	Ureinwohner	Lehrerschaft
Waffeleisen	Schwalben	Stall
Fortführung	Familienangehörige	Auftaktspiel

1.1 Nouns - Synonyms *(continued)*

Kampfhandlungen (*combat operations*)	**Rechner** (*computer*)	**Dürre** (*drought*)
Gefechte	Computer	Trockenheit
Hütchen	Rechenschaft*	Werktag
Abhandlungen*	Teerbau	Naturfreund
Spielhallen	Ausstoß	Schäferhund
Gefängnisse	Kontrabaß	Arthrose
Giraffen	Rückerstattung	Biene
Annäherungen	Gegenseite	Beratungszentrum
Spieße	Umschwung	Beat
Feuerwehrmänner	Ottomotor	Energie
Selbsttäuschung (*self-deception*)	**Ursprung** (*origin*)	**Freiraum** (*leeway*)
Selbstbetrug	Anfang	Spielraum
Selbstbedienung*	Seitensprung*	Innenraum*
Enttäuschung*	Urwald*	Freibier*
Kartoffelkäfer	Bordell	Wirt
Zuhause	Terminkalender	Zahlungsfähigkeit
Empfindung	Anhängsel	Bohème
Waldfriedhof	Zeichnung	Polizeichef
Landwirtschaft	Vormonat	Ostdeutscher
Schokoriegel	Kavallerie	Appetitzügler
Gradmesser (*indicator*)	**Leibesvisitation** (*body search*)	**Häme** (*malice*)
Maßstab	Durchsuchung	Schadenfreude
Hackmesser*	Leibesfrucht*	Pfund
Halbfinale	Wohltäterin	Lebenshilfe
Reiter	Zwischenzeit	Brustkorb
Kirchplatz	Nashorn	Schnellstraße
Überalterung	Gedenktag	Urteilsspruch
Zeitaufwand	Begleiterscheinung	Geisteszustand
Sexismus	Bekennerbrief	Bundesstraße
Überschwemmung	Rettung	Sammelband

1.1 Nouns - Synonyms *(continued)*

Nervenkitzel (*thrill*)	**Beischlaf** (*intercourse*)	**Seuche** (*disease*)
Risiko	Koitus	Krankheit
Nervengas*	Beiboot*	Anstand
Nervensystem*	Mittagsschlaf*	Wohnungsmarkt
Bürgerliste	Tiefschlaf*	Güte
Bösartigkeit	Homosexualität	Laus
Glied	Hörgerät	Nachbarstadt
Hörgerät	Überraschungsgast	Altersgruppe
Flosse	Sonderausstellung	Madonna
Gerichtsmedizin	Pfarrheim	Umgangssprache
Anmut (*charm*)	**Sachlage** (*circumstance*)	**Dirnen** (*prostitutes*)
Grazie	Sachverhalt	Huren
Großmut*	Sachbuch*	Birnen*
Freimut*	Auflage*	Situationen
Anbau*	Anlagenbau	Banken
Gatte	Aussaat	Erdnüsse
Ausdrucksweise	Choreografin	Sportereignisse
Legalität	Zutritt	Kapuzen
Pappkarton	Verkehrsführung	Cockpits
Wahlspruch	Wohnviertel	Beilagen
Ansteckungsgefahr (*risk of infection*)	**Theke** (*bar*)	**Semmeln** (*buns*)
Infektionsgefahr	Tresen	Brötchen
Hotelgewerbe	Apotheke*	Dackel
Verdunkelungsgefahr*	Arbeitsplatzverlust	Opas
Ausgleich	Eis	Zufälle
Bratwurst	Clinch	Sicherheiten
Biathlet	Treue	Kennerinnen
Verzweiflungstat	Chemieunterricht	Kumpels
Schlussfolgerung	Außenminister	Schöpfungen
Tierhaltung	Willkür	Marxisten

1.1 Nouns - Synonyms *(continued)*

Geschöpf (*creature*)	**Provenienz** (*origin*)	**Ausrottung** (*eradication*)
Lebewesen	Herkunft	Vernichtung
Abpfiff	Suppe	Ausbildung*
Weltmacht	Fladenbrot	Ausbeutung*
Urheberschaft	Papi	Wandertag
Bundesrat	Arbeitslosengeld	Nusskuchen
Speerwurf	Alterspräsident	Bildfläche
Stadtrand	Hack	Verwaltungsbezirk
Völkerkundemuseum	Kurpark	Grenzstein
Marktanalyse	Geburtagsgeschenk	Nahverkehr
Ansporn (*incentive*)	**Weltmacht** (*world power*)	**Klage** (*lawsuit*)
Anreiz	Großmacht	Anklage
Anbau*	Weltkugel*	Baufirma
Tötung	Ohnmacht*	Fehleinschätzung
Theateraufführung	Rettungsdienst	Autopsie
Volksvertretung	Marketing	Wertstoff
Athletik	Tänzer	Leiter
Kochlöffel	Mosambikaner	Rasse
Bombe	Trainingsplatz	Leserschaft
Guiness	Schläfrigkeit	Investition
Aggression (*aggression*)	**Eingebung** (*intuition*)	
Angriff	Intuition	
Aggregat*	Kundgebung*	
Depression*	Umgebung*	
Traumjob	Einziehung*	
Verpackung	Dolmetscher	
Ausweis	Kaserne	
Gabel	Vierbeiner	
Knebel	Ortsdurchfahrt	
Reif	Tierzüchter	

1.2 Nouns - Hypernyms/Hyperonyms

Lagerhalle (*storage hall*)	**Neuauflage** (*new edition*)	**Ausweichmöglichkeit** (*alternative*)
Halle (*hall*)	Auflage (*edition*)	Möglichkeit (*possibility*)
Küchenuhr	Neuankömmling*	Gesundheitsberatung
Fernstraße	Neubau*	Gastfamilie
Übergabe	Palmsonntag	Strafbefehl
Ampel	Unbeweglichkeit	Hauptstadt
Menetekel	Zukunftsperspektive	Nachlass
Bibliothek	Rauch	Überdruss
Musiker	Konsortium	Autohersteller
Autobahnabfahrt	Finanzverwaltung	Willen
Hornhaut (*horny skin*)	**Hof** (*court*)	**Volkswirtschaft** (*nat'l ecocomy*)
Haut (*skin*)	Fürstenhof (*royal court*)	Wirtschaft (*economy*)
Hornkonzert*	Zwangslage	Militärflughafen
Käsetorte	Zentralrat	Wirtshaus*
Tanzkapelle	Kreisel	Abart
Gründung	Betriebsamkeit	Saldo
Inhalt	Männchen	Gips
Journalismus	Abstellung	Wahlkampagne
Zunft	Streichinstrument	Ehrenmitglied
Torchance	Badearzt	Dauerregen
Bildschirm (*screen*)	**Grundton** (*keynote*)	**Dienstanweisung** (*job instr.*)
Großbildschirm (*big screen*)	Ton (*note*)	Anweisung (*instruction*)
Regenschirm*	Grunderwerb*	Kommunalisierung
Bildbericht*	Grundfeste*	Dienstbarkeit*
Mitwirkung	Dividende	Weltklasseschwimmer
Gipfelkonferenz	Übermaß	Weissagung
Geldspende	Weitschuss	Kampfflugzeug
Weizenkleie	Lebensgefährtin	Röhre
Sichtweise	Nebensatz	Rom
Clubmitglied	Redaktionsschluss	Kunstmesse

1.2 Nouns - Hypernyms/Hyperonyms *(continued)*

Amtsmissbrauch (*m. of authority*)	**Leuchtpistolen** (*signal p.*)	**Überschwemmung** (*flood*)
Missbrauch (*misuse*)	Pistolen (*pistols*)	Naturkatastrophe (*nat. des.*)
Amtsdeutsch*	Leuchtreklamen*	Überschneidung*
Amtshandlung*	Cassettenrecorder	Überschwang*
Schwert	Schwingungen	Befremdung
Kommunist	Jagdbomber	Benehmen
Beschleunigung	Motorhauben	Körperfunktion
Port	Baukörper	Sternstunde
Kicker	Anteilseigner	Wirtschaftsleistung
Schwefelsäure	Einsichten	Drogist
Hebebühne (*lifting platform*)	**Ramschbuden** (*junk shops*)	**Geschlecht** (*clan*)
Bühne (*platform*)	Buden (*shacks*)	Adelsgeschlecht (*nobility*)
Freilichtbühne*	Experten	Geschlechtsverkehr*
Ackerland	Urlaubsreisen	Schlagzeile
Verderb	Landeszentralbanken	Spendenaktion
Kapelle	Verkehrsämter	Beweisführung
Zahn	Mitmenschen	Artikel
Lippenstift	Teilnehmer	Staatsfeind
Zeitungsvereleger	Busunglücke	Stadtverwaltung
Crew	Mitbringsel	Mystik
Kronzeuge (*main witness*)		
Zeuge (*witness*)		
Kronleuchter*		
Kronjuwel*		
Kronkolonie*		
Parlamentspräsident		
Adel		
Hauptsache		
Kulturdezernentin		
Boden		

1.3 Nouns - Antonyms

Ungleichgewicht (*imbalance*)	**Gleichberechtigung** (*equality*)	**Entlastung** (*relief*)
Gleichgewicht (*balance*)	Diskriminierung (*discrimination*)	Belastung (*burden*)
Halbschwergewicht*	Gleichgültigkeit*	Entartung*
Federgewicht*	Gleichförmigkeit*	Entwarnung*
Katholizismus	Wesenszug	Kriegsschiff
Kuchenteller	Strahlung	Spanier
Aufwärmtraining	Blackout	Strampelhose
Lehrerschaft	Polizeibeamter	Stange
Stall	Platzwunde	Verweis
Auftaktspiel	Fabrik	Gentechnologie
Ossi (*East German*)	**Demut** (*humility*)	
Wessi (*West German*)	Hochmut (*arrogance*)	
Stimmabgabe	Armut*	
Bronzezeit	Unmut*	
Jugendmeister	Engländerin	
Auster	Bedarf	
Erdkruste	Hexenschuß	
Zuständigkeit	Aufmerksamkeit	
Form	Wichtigkeit	
Töpferkurs	Wissenschaftsmagazin	

2.1 Verbs - Synonyms

erniedrigen (*to humiliate*)	**zücken** (*to draw*)	**erwischen** (*to catch*)
demütigen	ziehen	ertappen
erübrigen*	abrücken*	wischen*
blitzen	entzücken*	auswischen*
plaudern	glücken*	wegwischen*
verrechnen	zucken	erwähnen
veröden	lächeln	klönen
schwenken	durchringen	runzeln
vorkämpfen	polieren	lieben
husten	ergehen	bereuen
verrichten (*to do a job*)	**traktieren** (*to maul*)	**versäumen** (*to miss*)
ausführen	quälen	verpassen
richten*	adaptieren*	säumen*
ausrichten*	relativieren	konsolidieren
errichten*	verstreichen	leiden
einrichten*	schwanken	versuchen
goutieren	anstacheln	einigen
eintauchen	abhauen	abspülen
verstopfen	querlegen	bohren
hinzufügen	beteuern	riskieren
verbannen (*to ban*)	**preisen** (*to praise*)	**wickeln** (*to wrap*)
abschieben	loben	einpacken
verleumden*	beibehalten	abwickeln*
weisen	abkratzen	entwickeln*
erzittern	abdrehen	anreizen
dirigieren	bemerken	klarmachen
quasseln	erhärten	aufrechterhalten
auftun	täuschen	schonen
lynchen	wiedergeben	erbarmen
lüften	antanzen	benehmen

2.1 Verbs - Synonyms *(continued)*

behütet (*protected*)	**spotten** (*to mock*)	**befugt** (*authorized*)
beschützt	lästern	ermächtigt
gesteckt	rotten*	dazugekommen
experimentiert	einmotten*	sympathisiert
beantwortet	reifen	rationiert
dramatisiert	beißen	votiert
gefüllt	offenlegen	gefeilscht
entschuldet	installieren	gekrümmt
probiert	adoptieren	deklariert
ausgeruht	plazieren	bekehrt

2.2 Verbs - Hyponyms/Hyperonyms

knabbern (*to nibble*)	**beheben** (*to remedy*)	**jammern** (*to moan*)
essen (*to eat*)	korrigieren (*to rectify*)	stöhnen (*to groan*)
kleckern*	heben*	klammern*
knacken*	erheben*	fragen
geben	entheben*	aufblähen
mieten	aufheben*	eskalieren
besehen	entgegensehen	hinbekommen
vermindern	mieten	einmischen
bröckeln	respektieren	gieren
entdecken	ausscheiden	durchschlagen

2.2 Verbs - Hyponyms/Hyperonyms *(continued)*

bergen (*to salvage*)	**orten** (*to locate*)	**karikieren** (*to caricature*)
retten (*to save*)	entdecken (*to discover*)	parodieren (*to parody*)
verschönern	horten*	markieren*
stanzen	wanken	riskieren*
rühren	überstrapazieren	hervorrufen
zutrauen	veranschaulichen	schwarzfahren
kriminalisieren	ticken	buckeln
satteln	navigieren	komplizieren
fortführen	benachrichtigen	nuckeln
untersagte (*prohibited*)	**lehren** (*to teach*)	**behindern** (*to impede*)
verbot (*forbade*)	dozieren (*to lecture*)	stören (*to disturb*)
versagte*	kehren*	verhindern*
unternahm*	auslegen	erhaschen
zusagte	manövrieren	vorspielen
herauskam	absenden	entschuldigen
ausschloss	absegnen	übelnehmen
erteilte	rühren	postulieren
dramatisierte	scharren	merken
ausschrieb	vergeuden	argwöhnen
schlendern (*to stroll*)	**beneide** (*envy*)	
gehen (*to go*)	missgönne (*to begrudge*)	
ändern*	beschneide*	
kränzen	schneide*	
bedanken	zwinge	
kursieren	teste	
zureden	übersehe	
drangsalieren	scheide	
schnauben	horte	
bestaunen	handle	

2.3 Verbs - Antonyms

erlauben (*to allow*)	**entschärft** (*defused*)
verbieten (*to forbid*)	verschärft (*to aggravate*)
beurlauben*	eingeschärft*
aufklauben*	entführt*
verschmerzen	gemalt
abfüllen	übertroffen
liebäugeln	ertappt
fristen	zurückgewiesen
interpretieren	aufgerissen
entleeren	getrennt

3.1 Adjectives - Synonyms

miese (*poor*)	**uneingeschränkte** (*unlimited*)	**verlässlich** (*reliable*)
üble	völlige	zuverlässig
langfristige	unbedachte*	unerlässlich*
beige	eigentliche	unbewusst
unhaltbare	unlösbare	unähnlich
logische	individuelle	säuerlich
einjährige	vertrauenserweckende	untätig
harte	separatistische	physisch
mitteleuropäische	verhasste	stur
grelle	rosige	bombastisch
innigen (*heartfelt*)	**gegenwärtige** (*current*)	**unerfreuliche** (*unpleasant*)
herzlichen	jetzige	missliche
inneren*	gegenläufige*	unerhörte*
unkontrollierbaren	gegenteilige*	unerreichbare*
apokalyptischen	gegensätzliche*	leidtragende
neuseeländischen	einflussreiche	verkehrsberuhigte
kosmetischen	doppelte	symphonische
längerfristigen	schriftliche	glasklare
musikbegeisterten	unbewohnte	dreistellige
kirchlichen	krebskranke	helle

3.1 Adjectives - Synonyms *(continued)*

mürbe (*brittle*)	**frühzeitige** (*early*)	**lächerlich** (*ridiculous*)
brüchig	frühe	lachhaft
müde*	kurzzeitige*	vertraulich*
barbarisch	gleichzeitige*	charakterlich*
lebenswichtig	zwischenmenschliche	geil
unhaltbar	auffällige	sittenwidrig
endgültig	holsteinische	fit
schrill	beachtliche	feinfühlig
ungeplant	alteingesessene	einsehbar
lebensfähig	heimliche	entsagungsvoll

3.2 Adjectives - Hyponyms/Hyperonyms

vermeintlichen (*alleged*)	**wackligen** (*wobbly*)	**fristlose** (*without notice*)
trügerischen (*deceptive*)	instabilen (*unstable*)	sofortige (*immediate*)
öffentlichen*	wackeren*	hemmungslose*
polizeilichen	ungläubigen	haltlose*
initimen	jüdischen	lyrische
geschmackvollen	adeligen	ahnungslose
ertragreichen	bildhaften	versehentliche
bürgerlichen	leibhaftigen	qualitative
unvergesslichen	maßgeblichen	gewaltsame
fürstlichen	unendlichen	dreifache
sentimental (*sentimental*)	**diffus** (*diffuse*)	
gefühlvoll (*soulful*)	unklar (*blurry*)	
mental*	einflussreich	
instrumental*	diffizil*	
fundamental*	aussagekräftig	
monumental*	strafrechtlich	
langwierig	erlogen	
spröde	zahlenmäßig	
partiell	aggressiv	
ukrainisch	geistig	

3.3 Adjectives - Antonyms

labil (*unstable*)	**unberechenbaren** (*incalculable*)	**zerstörerische** (*destructive*)
stabil (*stable*)	kalkulierbaren (*calculable*)	konstruktive (*constructive*)
mobil*	unberechtigten*	zerstreute*
trocken	unerreichbaren*	portugiesische
andächtig	glorreichen	vorurteilsfreie
infektiös	erbarmungslosen	nordöstliche
berufstätig	besitzlosen	unbegreifliche
relevant	perspektivischen	stoffliche
furchtbar	tonangebenden	kollegiale
betulich	südamerikanischen	haarsträubende
ledige (*unmarried*)	**verständlich** (*understandable*)	**reale** (*real*)
verheiratete (*married*)	unverständlich (*incomprehensible*)	irreale (*unreal*)
lederne*	verständig*	solare *
gnädige*	freundlich*	reizvolle
gesundheitliche	brüsk	sechstätige
metaphorische	analog	friedenserhaltende
gestalterische	wehrlos	humanistische
spielfreie	angenehm	millionenschwere
unbefangene	konditionell	verbale
verschnupfte	solide	vegetarische
rationale (*rational*)	**militärische** (*military*)	**lösbar** (*solvable*)
unvernünftige (*unreasonable*)	unverständlich (*incomprehensible*)	unlösbar (*unsolvable*)
rationelle*	militante*	genießbar*
nationale*	atmosphärische*	begehbar*
ideenlose	irrsinnige	nebligtrüb
hastige	sachdienliche	abartig
restriktive	haitianische	leibhaftig
liebenswürdige	wechselseitige	kauzig
nominale	wahnsinnige	sachgerecht
wacklige	rothaarige	unausgesprochen

3.3 Adjectives - Antonyms *(continued)*

formell (*formal*)	**begreiflich** (*understandable*)
informell (*informal*)	unverständlich (*incomprehensible*)
lautlos	begrifflich*
humoristisch	handgreiflich*
gewohnt	schief
kräftesparend	irreparabel
nebenberuflich	gegenwärtig
märchenhaft	modisch
reizvoll	missverständlich
parteiintern	schwanger

Appendix H

Human Word Similarity Ratings for the German Noun Pairs

This appendix contains the average native speaker ratings for word similarity of 57 noun pairs. The noun pairs are approximate translations of the respective items in the Rubenstein and Goodenough (1965) data set. The word pair containing nouns not included in any of the LSA word spaces is marked with an asterisk.

Zauberer - Magier (*wizard - magician*)	3.96
Leibeigener - Sklave* (*serf - slave*)	3.83
Edelstein - Juwel (*gem - jewel*)	3.83
Junge - Bursche (*boy - lad*)	3.79
Kraftfahrzeug - Auto (*automobile - car*)	3.79
Forst - Wald (*woodland - forest*)	3.75
Küste - Ufer (*coast - shore*)	3.67
Autogramm - Unterschrift (*autograph - signature*)	3.54
Mittag - Mittagsstunde (*midday - noon*)	3.54
Vogel - Kranich (*bird - crane*)	3.54
Hügel - Berg *hill - mountain*)	3.46
Backofen - Herd *furnace - stove*)	3.42
Grinsen - Lächeln (*grin - smile*)	3.38
Schnur -Seil *string - cord*)	3.38
Nahrung - Obst (*food - fruit*)	3.29
Glas - Becher (*glass - tumbler*)	3.25
Fahrt - Reise (*journey - voyage*)	3.25
Vogel - Hahn (*bird - cock*)	3.17

Bruder - Mönch (*brother - monk*)	3.04
Friedhof - Kirchhof (*cemetery - graveyard*)	3.00
Gerät - Werkzeug (*implement - tool*)	3.00
Auto - Fahrt (*car - journey*)	2.75
Kranich - Hahn (*crane - rooster*)	2.21
Kran - Werkzeug (*crane - implement*)	1.96
Nahrung - Hahn (*food - rooster*)	1.88
Berg - Wald (*mountain - forest*)	1.75
Zauberer - Orakel (*wizard - oracle*)	1.71
Berg - Küste (*mountain - coast*)	1.71
Vogel - Wald (*bird - forest*)	1.63
Bruder - Bursche (*brother - lad*)	1.58
Fabel - Magier ((*fable - magician*)	1.54
Küste - Reise (*shore - voyage*)	1.46
Ufer - Wald (*shore - forest*)	1.29
Ufer - Hügel (*shore - hill*)	1.25
Orakel - Fabel (*oracle - fable*)	1.25
Küste - Forst (*shore - woodland*)	1.08
Glas - Juwel (*glass - jewel*)	1.08
Backofen - Werkzeug (*furnace - implement*)	1.04
Friedhof - Wald (*cemetery - forest*)	0.96
Obst - Backofen (*fruit - furnace*)	0.92
Friedhof - Hügel (*cemetery - hill*)	0.92
Glas - Zauberer (*glass - wizard*)	0.58
Mönch - Sklave (*monk - slave*)	0.58
Bursche - Magier (*lad - magician*)	0.58
Grinsen - Bursche (*grin - lad*)	0.58
Mönch - Orakel (*monk - oracle*)	0.54
Forst - Kirchhof (*woodland - graveyard*)	0.46
Junge - Fabel (*boy - fable*)	0.38
Friedhof - Psychiatrie (*cemetery - asylum*)	0.38
Junge - Hahn (*boy - rooster*)	0.29

Kraftfahrzeug - Magier (*automobile - magician*)	0.04
Autogramm - Küste (*autograph - shore*)	0.04
Mittag - Schnur (*noon - string*)	0.04
Grinsen - Werkzeug (*grin - implement*)	0.00
Hahn - Reise (*rooster - voyage*)	0.00
Seil - Lächeln (*cord - smile*)	0.00
Berg - Herd (*mountain - stove*)	0.00

Appendix I

Sample Questionnaire for Evaluation Study

FRAGEBOGEN ZUR BEWERTUNG VON BEISPIELSÄTZEN

VIELEN DANK FÜR IHRE TEILNAHME AN DIESER STUDIE.

Der Zweck dieser Studie ist, mögliche Beispielsätze für schwierige bzw. unbekannte Wörter in einem Lesetext zu bewerten. Die Bewertung der Beispielsätze sollte unter dem Gesichtspunkt erfolgen, inwieweit sie geeignet sind, die Bedeutung des jeweiligen Wortes zu illustrieren. Zielgruppe für die Beispielsätze sind Deutschlernende mit Englisch als Muttersprache.

Es ist beabsichtigt, diese Beispielsätze zukünftig in einer computergestützten Lernumgebung anzuwenden, bei der die Studenten beim Lesen von deutschen Texten auf für sie schwierige oder unbekannte Wörter 'klicken' können, um unter diversen Hilfsoptionen auszuwählen (z.B. bildliche Illustrierungen, Übersetzungen, Definitionen oder eben Beispielsätze).

IHRE AUFGABE:

Ihnen werden 20 verschiedene Wörter präsentiert, die aus Zeitungs- oder Magazintexten ausgewählt wurden. Diese Wörter sind durch Fettdruck und Unterstreichung hervorgehoben und werden zusammen mit den dazugehörigen Satzauszügen gezeigt.

Nachdem Sie diese Sätze gelesen haben, bewerten Sie bitte 8-10 potentielle Beispielsätze für das jeweilige Wort danach, für wie hilfreich Sie die Beispiele erachten. Bei ambigen Wörtern (z. B. *vertreiben*) können Sie davon ausgehen, dass alle gezeigten Beispielsätze das Wort in etwa dem gleichen Sinn enthalten, in dem es im Originalsatz erscheint. Bitte bewerten Sie die Beispielsätze auf einer Skala von 1 (nicht hilfreich) bis 9 (sehr hilfreich); bitte nutzen Sie bei Ihrer Bewertung die volle Bandbreite der Skala soweit wie möglich.

Bitte berücksichtigen Sie bei Ihrer Bewertung, dass die Muttersprache der Studenten Englisch ist. Da die Zielworte in ihrem Schwierigkeitsgrad variieren, sollte kein bestimmter Kenntnisstand der Schüler angenommen werden. Sie können jedoch voraussetzen, dass die Studenten in ihren Deutschkenntnissen soweit fortgeschritten sind, dass Grammatik oder schwierige Satzsyntax für sie kein Problem darstellen. Bitte beachten Sie, dass einige Beispielsätze ohne den dazugehörigen Kontext schwer verständlich erscheinen mögen. Sie werden gebeten, Ihre Bewertung nicht von diesem Umstand beeinflussen zu lassen.

IHR NAME:

EXAMPLE SENTENCES EVALUATION QUESTIONNAIRE

THANK YOU FOR PARTICIPATING IN THIS STUDY.

The purpose of the study is to test for the most helpful example sentences for unknown or difficult target words in reading texts. The example sentences should be judged according to how helpful they are in illustrating the meaning of the respective target word. The target audience for these examples are students of German as a Foreign Language.

The intended application for the example sentences to be used is a computer-assisted reading environment, where students reading German texts can click on unknown or difficult words and choose among different explanation options, such as pictorial glosses, translations, definitions, or example sentences.

YOUR TASK:

You will be presented with 20 different words which are taken from different newspaper or magazine articles. The words (in boldface and underlined) are presented together with their respective sentences.

After reading the sentences, you will be asked to rate 8-10 potential example sentences for each word according to how helpful you consider them. In the case of ambiguous words (e.g. *vertreiben*), you may assume that the example sentences shown contain the word in roughly the same sense as in the original sentence. To this end you will be provided with a scale from 1 (not helpful) to 9 (very helpful) on which to rate the helpfulness of each example sentence. Please use the full range of the rating scale as far as possible.

When rating the examples, please bear in mind not only the original sentence but also that the students' native language is English. Since the target words tend to vary in their level of difficulty, no particular level of proficiency is assumed for the students. You may assume that the students are sufficiently advanced learners for whom grammar and potentially difficult syntax of reading texts are not a problem. Beware that some example sentences may appear difficult to understand because they are presented out of context. You are asked to not let this aspect influence your ratings of the sentences.

WORT 1: Gebäck

Um 610 herum, so viel weiß die Wissenschaft immerhin, formte er den Teig, der beim Brotbacken übrig geblieben war, zu einem **Gebäck**, das aussehen sollte wie betende Kinderhände.

Bsp 1: Beim Backen dieses Gestalten- oder Gebildegebäcks greift er auf alte, überlieferte Rezepte zurück und benutzt zum Ausstechen des Gebäcks Formen seiner Vorfahren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Um so mehr schmeckte nach der Schule so ein Gebäck.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Mit dabei die Produzenten des kalorienträchtigen Gebäcks, die Bäckerleute Ingrid und Uwe Richter vom Boulevard, und der sechsjährige Michael.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Es gibt Glühwein, andere Getränke, Kaffee und Kuchen, Gebäck, Waffeln und Gänsebraten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Aber immerhin kann man Gebäck mit dem Pulver aromatisieren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Nach dem Gottesdienst sind alle zu Gebäck, Getränk und Gespräch eingeladen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Hier duftet es nach frischem Gebäck, dort nach Gebratenem.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Wenn man in Deutschland ist, sollte man unbedingt das dort typische Gebäck essen, wie z.B. Käsebrötchen, Brezel, Mohnbrötchen, Schusterjungen usw.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Sie servierte Gebäck zum Kaffee.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 2: Currywurst

Currywurst kann man in Hua Hin essen, Kassler mit Sauerkraut ist auch kein Problem, und Papa Joe, ein Berliner, macht die besten Steaks.

Bsp 1: Currywurst rot-weiss, wie es hier so schön heisst.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Dem Ministerpräsidenten mit Kanzlerambitionen wiederum geht es um die Currywurst und das Steak, die ihm seine vegetarische Ehefrau vorenthalten habe.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Eigenkapitalrendite und Dividenden, schön und gut, was diesen Schlag Aktionär wirklich interessiert, sind Naturaldividenden wie Currywurst, Putenbrust und Präsente zum Mitnehmen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Obwohl es vielleicht merkwürdig klingt, ist Currywurst eine Ur-Berliner Spezialität.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Die Berliner Erfinderin der Currywurst, Herta Heuwers, ist tot.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Man verspeiste mehrere Currywürste.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Oder, letztes Beispiel: die Currywurst.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Nicht immer nur Currywurst.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 3: durchaus

Das war beim Oder-Hochwasser 1997 so, und auch die Kölner Altstadtbewohner kennen diese **durchaus** unwillkommenen Besucher, die so regelmässig in der Domstadt erscheinen, wie der Rhein über die Ufer schwappt.

Bsp 1: Ein sehr gutes Rennrad kann durchaus 5000 Euro kosten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Dies könnte durchaus nach 1996 der Fall sein.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Jetzt soll die neue Schule davon Zeugnis ablegen, dass die Farbe durchaus in das Stadtbild passt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Die Marke ist zwar positiv besetzt, hat aber durchaus auch Runzeln und Fältchen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Das Geschäft mit der Bundesliga lasse sich durchaus profitabel betreiben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Die Gewitter könnten durchaus recht heftig sein.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Ich bin nicht bereit, Ihre Argumente ohne Weiteres zu akzeptieren, bin aber durchaus bereit, sie mir anzuhören.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Den derzeitigen Kölner Skandal deckt diese Definition durchaus ab.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Sie hat auch keine Anzeichen dafür, dass die Ganoven des Rhein-Main-Gebietes ihre Tatort-Grenzen fließend halten: "Wir haben durchaus eine hausgemachte Kriminalität.".

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: Es hat sich in der Vergangenheit mehr als einmal gezeigt, dass die Macht der USA durchaus an Grenzen stößt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 4: reagieren

Geradezu genüßlich haben konservative Politiker und anti-europäische Zeitungen der Insel auf die neue Misere in Frankreich und auf den teutonischen Schock über das Ende der Selbsttäuschung in Deutschland **reagiert.**

Bsp 1: Darauf müsse die Politik reagieren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Es gehe darum, auf die Liberalisierung im Bahnverkehr in Europa zu reagieren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Aufstiegsorientierte ausländische Familien dagegen reagieren längst: Sie melden ihre Kinder in Schulen mit niedrigem Ausländeranteil an.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: «Microsoft» hat auf den NC mit einem «Netz-PC» reagiert, einem abgespeckten PC, dessen Betriebskosten durch zentral gesteuerte Wartungsprogramme gesenkt werden sollen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Wie reagiere ich denn in Situationen, wenn das Kind permanent schreit, ist es krank oder was hat es denn?

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Unfähig, zu reagieren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Zur Pause reagierte der französische Teamchef Michel Platini.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Wir haben versucht, unseren Lehrer durch unser schlechtes Benehmen zu provozieren, er hat aber darauf überhaupt nicht reagiert.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Er schien auf ihre Worte kaum zu reagieren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: Und schon reagieren die Anwohner nach dem Motto: "Schafft diese Menschen irgendwohin, nur nicht in unsere Nähe".

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 5: allerlei

Preussen war auf dem Weg zur europäischen Großmacht und legte sich im Laufe der Jahre **allerlei** kulturelle Bauten zu.

Bsp 1: Das Magazin DM hat sich in seiner Online-Version (http://www.DM-online.de) allerlei einfallen lassen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Für diese Zeit hat sich Andreas Juchli deshalb allerlei vorgenommen: "Ich möchte ein Buch schreiben", verrät er.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Musik, allerlei "Überraschungen" und natürlich reichlich Essen und Getränke werden beim Sommerfest des Jugendtreffs am Freitag, 24. Juli, geboten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Man hört so allerlei.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Immer wieder rennen die Kinder weg, suchen nach schönen Steinen, Larven, Pflanzen und allerlei Getier.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Ohne Glitzer scheint es bei Chanel nicht zu gehen: Viele der weichen Wollstoffe sind mit Pailletter oder allerlei Klimperzeug geschmückt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Dazu gibt's eine Weindegustation aus Yvonnes Weinkabinett und allerlei, vom Buch inspirierte Köstlichkeiten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Wen der Weg hungrig gemacht hat, kann unter allerlei mongolischen und nichtmongolischen Speisen und Getränken wählen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Diana hat allerlei bunte Klamotten spottbillig auf dem Flohmarkt ergattert.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 6: Einheit

Von denen ist nach zwölf Jahren **Einheit** nicht mehr viel zu lernen.

Bsp 1: Viele Ostdeutsche glauben, dass die sogenannte deutsche Einheit 2006 eine Illusion ist.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Deutliche Unterschiede zwischen Ost und West haben sich auch nach zwölf Jahren deutscher Einheit im naturwissenschaftlichen Unterricht gehalten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Körper, Seele und Geist machen die Einheit Mensch aus.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Im dritten Teil der Ausstellung werden Briefe gezeigt, die sich mit der Deutschen Einheit befassen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Man sieht, die Mannschaft ist eine Einheit, die gut auf dem Platz funktioniert.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Im Friedensabkommen von Taif 1990 wurde bestimmt, dass es eine Regierung der nationalen Einheit geben werde: Das geschah nicht.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Die drei Stücke bilden eine Einheit.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Er erkennt die Grundsätze des Roten Kreuzes, welche lauten: Menschlichkeit, Unparteilichkeit, Neutralität, Unabhängigkeit, Freiwilligkeit, Einheit, Universalität.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 7: Meinungsumfrage

In einer **Meinungsumfrage** sprachen sich lediglich 35 Prozent der Befragten für eine britische Beteiligung an einer Neuauflage des Golfkriegs aus.

Bsp 1: In Deutschland werden angeblich jedes Jahr 1,3 Milliarden Mark für Meinungsumfragen zu allen möglichen und unmöglichen Themen ausgegeben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: In einer Meinungsumfrage in Großbritannien hat man festgestellt, dass über 50% der Befragten nicht rauchen und nie geraucht haben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Rund 60 Prozent befürchten, mit den islamistischen Terroristen in einen Topf geworfen zu werden, fand eine Meinungsumfrage heraus.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Die Meinungsumfragen zeigen uns zwar eine überwältigende Ablehnung.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Zur allgemeinen Überraschung und entgegen den meisten Meinungsumfragen ist aber die DK auch bei den Parlamentswahlen beträchtlich hinter der Partei Iliescus zurückgeblieben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Die jüngsten Meinungsumfragen geben Labour 19 Prozent Vorsprung vor den Konservativen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Laut der neuesten Meinungsumfrage sind 60% der Bevölkerung für das neue Gesetz.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Sie hatten einen Tag zuvor die Ergebnisse einer Meinungsumfrage über das Sozialverhalten der Chinesen veröffentlicht.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Meinungsumfragen in Politik umzusetzen, scheint mir die einzige Begründung zu sein.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: In der jüngsten Meinungsumfrage erfuhr Putin die Zustimmung von 59 Prozent der Befragten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 8: Gesprächspartner

Tatsache ist, dass Arafat als Repräsentant der Palästinenser von den USA und der EU als **Gesprächspartner** akzeptiert wird.

Bsp 1: Als ich Skype zum ersten Mal benutzt habe, hatte ich Probleme meinen Gesprächspartner richtig zu hören.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Der Frankfurter hat gewiss nicht alle Gesprächspartner in Israel überzeugt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Er fand in ihr eine anregende Gesprächspartnerin.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Dies zur Vorstellung von Wolfgang Behrendt, mein Gesprächspartner jetzt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Wer Alfred Brendel gegenübersitzt, hat einen Gesprächspartner, der jedes Wort so ernst nimmt, als hinge die Welt davon ab.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Zu ausländischen Gesprächspartnern verbindet bisher nur die Telegate, allerdings nur in die USA, nach Großbritannien, Frankreich, Österreich und in die Schweiz.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Doch in Berlin sagte der Gast den Gesprächspartnern, er werde sich auf solch ein Tauschgeschäft nicht einlassen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Gesprächspartner werden unter anderen Ministerpräsident Kiichi Miyazawa und Kaiser Akihito sein.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Als Gesprächspartner sind auch Flüchtlinge aus Syrien, Algerien und Zaire zur Stelle.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: Selbstverständlich gibt Ihnen das System Auskunft darüber, wie Sie das Büro Ihres Gesprächspartners vom Flughafen aus am preiswertesten, schnellsten oder bequemsten erreichen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 9: Machthaber

Ausgerechnet dieser Stolz der kommunistischen **Machthaber**, für dessen Bau die Parteimitglieder freiwillig Geld spendeten, wurde Anfang der neunziger Jahre zu einem „Bank- und Finanzzentrum" umfunktioniert, dessen Hauptteil die neu entstandene Warschauer Börse bildete.

Bsp 1: Doch der Mann, den die Weltpresse mittlerweile zum geheimen Machthaber der Autonomiegebiete proklamiert hat und den angeblich sogar Arafat fürchtet, heißt Marwan Barghouti.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Die Machthaber in der DDR konnten es sich gar nicht anders erklären.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Anfang Januar hat die "Volksfront" des neuen Machthabers Blaise Compaores 1.500 Delegierte zur kritischen Bilanz der Sankara-Jahre versammelt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Obwohl die Iraker bereits eine Regierung gewählt haben, sind ohne Zweifel die Amerikaner und Briten noch die echten Machthaber im Lande.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Qualitätsjournalismus zeichne sich auch durch seine Unabhängigkeit gegenüber den Machthabern aus, sagte Zimmermann.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Als man Burundi in die Demokratie entließ, glaubten die politischen Machthaber, es sei damit getan, alle Ämter neu zu besetzen, erzählt Oskar Dür.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Aber genau das ist in vielen Staaten Afrikas nicht Priorität der Machthaber.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Viele Menschen spekulierten in diesen Stunden über die Zukunft des gestürzten Machthabers.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Die Machthaber dieses Landes waren nicht zu Verhandlungen bereit.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 10: Falle

Das führt sie in die andere **Falle**, die der Verwechselbarkeit.

Bsp 1: Heuer wurden im Burgenland bereits zwei Menschen durch Fallen verletzt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: In diese Falle war M. getappt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Schon mit meinem ersten Satz bin ich in die Falle gegangen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Mit einer riesigen Falle wollte er sie fangen, hatte aber keinen Erfolg.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Wir haben gedacht, dass unsere Gegner es mit uns ehrlich meinten, das war aber eine Falle: Ihr eigentliches Ziel war was ganz Anderes.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Das kann zur Falle werden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: "Heute werden sie mit dem synthetisch nachgemachten Duftstoff der Weibchen in Fallen gelockt", berichtet Professor Hans-Jürgen Otto vom Landwirtschaftsministerium in Hannover.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Für einkommensschwache Mieter ist Wohngeld der beste Ausgleich; sozialer Wohnungsbau darf nicht in die unsoziale Falle der Fehlbelegung führen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Die Polizei hat dem Dieb eine Falle gestellt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: Einmal eine schöne Lovestory ... aber ist das nicht eine Falle?

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 11: Todestag

Fast alle werden rund um ihren **Todestag** am 29. Mai auf vielen TV-Kanälen wiederholt.

Bsp 1: Der Zeitpunkt der Veröffentlichung überrascht allerdings – wäre Marleys zehnter Todestag im vergangenen Jahr doch ein besserer Anlass gewesen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Zum 30. Todestag hat der Club der Grinzing-Freunde einen Ehrengulden aufgelegt und Hans Moser damit "versilbert".

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Wir haben dieses Jahr den 200. Todestag von Mozart gefeiert.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Am 10. Dezember, Nobels Todestag, wird er zum 100. Mal verliehen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Alle Jahre wieder: Am 20. Februar gedenkt Tirol dem Todestag des Freiheitskämpfers Andreas Hofer.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Der 250. Todestag von Johann Sebastian Bach wurde feierlich begangen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Dies hat Oberbürgermeister Ulrich Bauer am Sonntag bei der Feierstunde aus Anlass des 50. Todestags Häckers bekanntgegeben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Alle waren sich einig: Es macht nur Sinn, wenn die Projekte genau am sechzigsten Todestag von Sophie und Hans Scholl vorgestellt würden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Hans-Jörg Neuschäfer schreibt zum 100. Todestag des spanischen Autors Clarin.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 12: vertreiben

Zwar wurden die Taliban, die Washington einst mit an die Macht gebracht hat, nun durch Washingtons Bomben von der Macht **vertrieben**, Kabul hat Chancen auf eine bessere Zukunft.

Bsp 1: Die Bundespolizei hatte am 25. Februar die Operation "Freier Urwald" gestartet, durch die die Goldsucher aus dem Yanomani-Reservat vertrieben werden sollen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Wiesenthal: "Die armen Juden wurden vertrieben, die reichen verbrannt – weil die Kirche ja kein Blut vergießen durfte."

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Uns gehen die jungen Hooligans in den Einkaufszentren auf die Nerven, und wir wollen sie (wenn's geht) von dort vertreiben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Kroatische Angreifer vertrieben moslemische Familien aus drei Dörfern.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Zwischen 1938 und 1944 wurden in Europa 15 Millionen Menschen aus dem mittelost- und südosteuropäischen Raum Richtung Osten vertrieben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Das vertreibt die Zuschauer.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Vertreib doch mal die Wespen!

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Ohne die Duldung Pakistans, dem Nachbarland Afghanistans wäre das Ziel, die Taliban von der Macht in Kabul zu vertreiben, in weite Ferne gerückt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 13: derzeitig

Die **derzeitigen** Probleme des Nationalteams hätten nichts mit der Person Klinsmann zu tun, sondern resultierten aus den 90er Jahren.

Bsp 1: "Gesund sein, heisst auch glücklich sein", meinte Graff, "Das derzeitige Gesundheitssystem wird dem nicht gerecht."

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Hatten Sie bei der derzeitigen politischen Lage Bedenken nach Österreich zu kommen?

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Nach dem derzeitigen Stand werden in Niederösterreich in den nächsten Jahren rund 6000 Personen für den Pflegebereich fehlen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Wie beurteilen Sie das derzeitige Klima in der Partei?

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Im derzeitigen Parlament sind die Schwarzen nicht vertreten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Die Bewerber könnten davon ausgehen, mit den "derzeitigen Verpflichtungen der KirchMedia" nichts mehr zu tun zu haben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Insgesamt sind im Zuge der derzeitigen Feuerserie in Colorado bis zu 80 Wohnhäuser abgebrannt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Seine derzeitige Freundin arbeitet bei der Zeitung.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Der derzeitige Koalitionspartner Bündnis 90/Die Grünen kommt auf sechs, die PDS erhält fünf Prozent der Stimmen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: Das derzeitige Wetter in Schottland lässt vermuten, dass es nie Frühling wird.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 14: neidisch

Freunde haben mir ein Publikum Spalier stehender, einsteigewilliger Damen sowie **neidisch** blickender, eingesperrter Zellenfahrer in Aussicht gestellt.

Bsp 1: Nun ist er neidisch auf mich.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Fast neidisch verfolgte ich fortan via Berichterstattung die Erlebnisse meiner Kolleginnen und Kollegen, die nach mir an der Reihe waren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Er ist neidisch auf ihre Jugend.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: So neidisch kann der Himmel sein?

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Anfangs gab es deswegen innerhalb der Klasse kleinere (neidische) Reibereien.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Das alles bleibt natürlich nicht ohne Konsequenz für Ihre Lebensführung und die neidischen Blicke mancher Zeitgenossen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Gestehe, dass ich neidisch bin.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Die ganze Republik soll neidisch werden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Wer schaut an einem lauen Sommerabend nicht einmal neidisch einem schnittigen Cabrio nach.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 10: Obwohl sein Schulfreund in der Zwischenzeit steinreich geworden war, war Johannes nicht neidisch auf ihn, da dessen Gesundheit nicht die beste war.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 15: Wassertemperatur

Die entlang dem Äquator verteilten Messbojen der Meteorologen haben in den vergangenen Wochen steigende **Wassertemperaturen** im westlichen und zentralen Pazifik registriert.

Bsp 1: Auch die Schwimmer im Hallenbad müssen sich auf kältere Zeiten gefasst machen: Die Wassertemperatur soll um ein Grad vermindert werden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Die Wassertemperaturen erreichen an der Adria 20, in der Ägäis 22 C.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: 566 Petitionäre möchten, dass der Wassertemperatur bei der umfangreichen Sanierung des Schwimmbades gebührende Beachtung geschenkt werde.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Die Wassertemperaturen des Mittelmeeres schwanken zwischen 22 und 24 C, wobei es wie immer im östlichen Mittelmeer wärmer als im westlichen ist.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Die augenblickliche Wassertemperatur der Nordsee beträgt 14 Grad Celsius.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Die Wassertemperaturen im westlichen Mittelmeer betragen 17 bis 21 C, im östlichen Mittelmeer 21 bis 23 C.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Noch lädt die Wassertemperatur im Schwimmbad, gestern Dienstag wurden noch kühle 17 Grad gemessen, nicht gerade zum Baden ein.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Ich werde die Wassertemperatur im Swimming Pool 1957 in Tirol nie vergessen – kalt!!

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Sechs Grad Celsius war die Wassertemperatur.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 16: offenbaren

Viele schleppten ihre Krankheit schon länger mit sich herum, **offenbarten** sich aber erst angesichts des Terrors ihren Ärzten, sagt Professor Klaus Wahle, der die Studie leitete.

Bsp 1: Nicht nur, dass sich Kinder kaum so vielen Menschen offenbaren.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Zunächst aber galt es, dieses zu prägen und dabei offenbarte sich wiederum sein Hang zur "Complete Organisation".

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Er hat immer versucht, seine panische Angst vor Spinnen zu verheimlichen, musste aber seine Angst offenbaren, als er in Australien mit einer Riesenspinne in seinem Bett konfontriert war.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Er hatte geschworen, nichts zu verraten, und ahnte nicht, dass er mir mit diesen Worten alles offenbart hatte.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Der rein quantitative Bundesländervergleich (siehe Grafik) offenbart eine äußerst unterschiedliche Handhabung der Sportförderung.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Er offenbarte sich seinem Freund.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Beginn des Konzertes, bei dem sich "die russische Seele offenbart", ist um 19.30 Uhr.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Die direkte Bezugnahme auf Bin Laden in dem Memorandum sei vorher von der US-Regierung nicht offenbart worden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: "Diese Daten offenbaren den bislang besten biologischen Hinweis auf einen klimatischen Wandel", schreiben die beiden Wissenschaftler in "Science".

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 17: Jauche

Andererseits hat sie dafür zehn Jahre lang Kübel voll **Jauche** über sich ausschütten lassen müssen.

Bsp 1: Vom Regen also in die Jauche?

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Ein Liter Wein in einem Fass Jauche gibt ein Fass Jauche.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Oder die durch alle Medien verbreitete Androhung, "Demonstranten mit Jauche" zu bekämpfen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Obwohl die Luft auf dem Lande im Allgemeinen schön riecht, kann man nicht immer den unangenehmen Geruch von Jauche vermeiden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Die Jauche zerstörte nämlich die Arbeit einer ganzen Woche.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Dünger wie Jauche, Gülle und Kompost sind organische Stickstoffdünger.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Im Osttiroler Defereggental stank im April der Moosbach nach Jauche.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Das stinkt wie Jauche.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 18: Innenstadtbereich

Lediglich im **Innenstadtbereich** und zum Flughafen wurde ein Notbetrieb aufrechterhalten.

Bsp 1: Insbesondere im Innenstadtbereich fehle es an radgerechten Wegen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Im Innenstadtbereich habe es jedoch keine entsprechenden Räume gegeben.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Wir hätten gerne im Innenstadtbereich oder auch im Ostend etwas genommen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Ebenfalls geknickt wurde der Plan, "bestimmte Innenstadtbereiche" von Bettlern oder anderen Umsatzschädlingen "freizuhalten".

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Es reicht lediglich für die Innenstadtbereiche.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Der Investorendruck auf den Innenstadtbereich steht der allseits geforderten Behutsamkeit wohl eher entgegen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Dem Mann werden zwölf Banküberfälle, die er seit 1998 vornehmlich im Innenstadtbereich begangen hat, zur Last gelegt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Viele deutsche Städte haben den Innenstadtbereich zur Fußgängerzone erklärt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 19: enttarnen

Staatsdiener, die in Bayern und Berlin als „Sex-Surfer“ **enttarnt** wurden, kamen mit einer Verwarnung davon.

Bsp 1: Ein Wahlkampfplakat der NPD, das die Förderkampagne für ein Holocaust-Mahnmal in Berlin verunglimpft, stammt aus der Feder des enttarnten V-Mannes Udo Holtmann.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Freddy Gut, Robi Meili, Susi Schläfli und die anderen werden ebenfalls als ohnmächtige Büttel des Zeitgeistes enttarnt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Die niederländische Verkehrsministerin Hanja Maij-Weggen, die bei den Autofahrern des Landes als meistgehasste Frau gilt, ist jetzt selbst als "Verkehrssünderin" enttarnt worden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Sie enttarnen sich gern selbst, weil sie im Verlauf des Gesprächs zu lachen beginnen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Hier werde also keine enttarnt, wie vielleicht im Restaurant eines teuren Hotels.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Max Stadler (FDP) unterstrich, V-Leute dürften keine Straftaten begehen, es sei denn, sie würden sich selbst enttarnen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: In Radebeul im Süden von Dresden wurde der Spion jetzt enttarnt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Johann hat gedacht, dass sein Interesse für Fußball ein Geheimnis war, aber er hat sich selbst enttarnt, als er sich während des ersten WM-Spiels krank gemeldet hat.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 9: Der Spion wurde schließlich doch enttarnt.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

WORT 20: eingehend

Die Aufzeichnungen belegen, dass Experten auf lokaler und Bundesebene die Bedrohungen durch den Hurrikan **eingehend** diskutiert hatten.

Bsp 1: Dabei haben die Betreuer Gelegenheit, eingehend mit den einzelnen Mädchen und Jungen zu reden, ihre Probleme kennenzulernen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 2: Er sprach eingehend mit Ganswind und Hermione.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 3: Obwohl man die potenziellen negativen Folgen eines Invasionskrieges im Irak eingehend erörtert hatte, haben die Koalitionspartner trotzdem für den Krieg entschieden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 4: Diese Frage wurde von Experten eingehend geprüft.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 5: Am Strand hatten meine liebe Frau und ich Musse, sie eingehend zu beobachten.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 6: Bahnreisende müssen sich eingehenden Kontrollen unterziehen, überall sind Wachleute unterwegs.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 7: Dieses Thema sei eingehend diskutiert worden.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Bsp 8: Aus der Präsidentschaftskanzlei verlautete, dass Bundespräsident Thomas Klestil ebenfalls beabsichtige, den endgültigen Textvorschlag von Experten eingehend prüfen zu lassen.

Nicht hilfreich 1 2 3 4 5 6 7 8 9 **Sehr hilfreich**

Appendix J

Published Papers

The following paper has been published by the author during the research leading up to this dissertation.

Segler, T., Pain, H. and Sorace, A. (2002). Second Language Vocabulary Acquisition and Learning Strategies in ICALL Environments. *Computer Assisted Language Learning*, 15/4: 409-422.

Bibliography

Abel, A. (2000). Das lexikographische Beispiel in der L2-Lexikographie (am Beispiel eines L2-Kontext- und Grundwortschatzwörterbuches). *Deutsch als Fremdsprache*, 37/3:163–169.

Aitchison, J. (1994). *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell, Oxford, 2nd edition.

Alderson, J. (2000). *Assessing Reading*. Cambridge University Press, Cambridge.

Antor, H. (1994). Strategien der Benutzerfreundlichkeit im modernen EFL-Wörterbuch. *Fremdsprachen Lehren und Lernen*, 23:65–83.

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 805–810.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Tannen, D., editor, *Coherence in Spoken and Written Discourse*, pages 45–80. Abilex, Norwood, NJ.

Beheydt, L. (1987). The semantization of vocabulary in foreign language learning. *System*, 15(1):55–67.

Beheydt, L. (1990). CALL and vocabulary acquisition in Dutch. In Kingston, P., Zähner, C., and Beutner, A., editors, *Languages, Continuity, Opportunity*, pages 186–192. CILT, London.

Béjoint, H. (1981). The foreign student's use of monolingual English dictionaries: A study of language needs and reference skills. *Applied Linguistics*, 2(3):207–221.

Black, A. (1991). On-line consultation of definitions and examples: Implications for the design of interactive dictionaries. *Applied Cognitive Psychology*, 5:149–166.

Bogaards, P. (1996). Dictionaries for learners of English. *International Journal of Lexicography*, 9(4):277–320.

Botel, M. and Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effort. *Elementary English*, 49:513–516.

Bowerman, B. and O'Connell, R. (1990). *Linear Statistical Models: An Applied Approach*. Duxbury, Belmont, CA, 2nd edition.

Brettschneider, G. (1978). *Koordination und Syntaktische Komplexität*. Wilhelm Fink Verlag, Munich.

Budanitsky, A. (1999). Lexical semantic relatedness and its application in natural language processing. Technical Report CSRG-390.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of ACL Worksop on WordNet and Other Lexical Resources*, pages 29–34, Pittsburgh.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32 (1):13–47.

Chaudron, C. (1982). Vocabulary elaboration in teachers' speech to L2 learners. *Studies in Second Language Acquisition*, 4:170–180.

Chaudron, C. (1983). Foreigner talk in the classroom — an aid to learning? In Seliger, H. and Long, M., editors, *Classroom-oriented Research in Language Acquisition*, pages 127–145. Newbury House, Rowley, MA.

Clahsen, H. and Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, 27:3–42.

Coady, J., Magoto, J., Hubbard, P., Graney, J., and Mokhtari, K. (1993). High frequency vocabulary and reading proficiency in ESL readers. In Huckin, T., Haynes, M., and Coady, J., editors, *Second Language Reading and Vocabulary Learning*, pages 217–228. Ablex Publishing Corporation, Norwood, NJ.

Collins (1991). *The Collins German Dictionary*. HarperCollins, Glasgow, 2nd edition.

Cook, G. (2001). 'The philosopher pulled the lower jaw of the hen.' Ludicrous Invented Sentences in Language Teaching. *Applied Linguistics*, 22/3:366–387.

Cook, R. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, New York.

Cowie, A. (1980). English dictionaries for the foreign learner. Paper presented at the Exeter Summer School on Lexicography.

Cowie, A. (1983). The pedagogical/learner's dictionary: English dictionaries for the foreign learner. pages 135–144. Academic Press, London.

Cowie, A. (1989). The language of examples in English learners' dictionaries. In James, G., editor, *Lexicographers and Their Works*, pages 55–65. University of Exeter, Exeter.

Cowie, A. (1999). *English Dictionaries for Foreign Learners: A History*. Oxford University Press, Oxford.

Creamer, T. (1987). Beyond the definition: some problems with examples in recent Chinese-English and English-Chinese bilingual dictionaries. In Cowie, A., editor, *The Dictionary and the Language Learner*, Lexicographica: Series maior ; 17, pages 238–245. Niemeyer, Tübingen.

Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.

Dagan, I., Lee, L., and Pereira, P. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 56–63.

De Florio-Hansen, I. (1994). *Vom Reden über Wörter: Vokabelerklärungen im Italienischunterricht mit Erwachsenen*. Narr, Tübingen.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science*, 21:391–407.

Drysdale, P. (1987). The role of examples in a learner's dictionary. In *The Dictionary and the Language Learner*, Lexicographica: Series maior ; 17, pages 213–223. Niemeyer, Tübingen.

DUDEN (2002). *Das Bedeutungswörterbuch*. Dudenverlag, Mannheim, 3rd edition.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press, Oxford.

Fellbaum, C. (1998). A semantic network of English verbs. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, pages 69–104. MIT Press, Cambridge, MA.

Field, A. (2005). *Discovering Statistics Using SPSS*. SAGE Publications, London, 2nd edition.

Firth, J. (1957). A synopsis of linguistic theory. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford. Special volume of the Philological Society.

Fodor, J., Bever, T., and Garrett, M. (1974). *The Psychology Of Language: An Introduction To Psycholinguistics And Generative Grammar*. McGraw-Hill, New York.

Fodor, J. and Garrett, M. (1967). Some syntactic determinants of sentential complexity. *Perception & Psychophysics*, 2 (7):289–296.

Foltz, P., Kintsch, W., and Landauer, T. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 2 & 3:285–307.

Fox, G. (1987). The case for examples. In Sinclair, J., editor, *Looking Up. An Account of the COBUILD Project in Lexical Computing*, pages 137–149. Collins, London.

Frazier, L. (1985). Syntactic complexity. In Dowty, D., Karttunen, L., and Zwicky, A., editors, *Natural Language Parsing*, pages 129–189. Cambridge University Press, Cambridge.

Frazier, L. (1988). The study of linguistic complexity. In Davison, A. and Green, G., editors, *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, pages 193–219. Lawrence Erlbaum, Hillsdale, N.J.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Gibson, E. and Warren, T. (1998). Discourse reference and syntactic complexity. Manuscript, MIT.

Goldstone, R. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):3–28.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer, Boston.

Grefenstette, G. and Nioche, J. (2000). Estimation of English and non-English language uses on the WWW. In *Proceedings of RIAO Conference on Content-based Multimedia Information Access*, pages 237–246, Paris, France.

Grewendorf, G. (1993). German: A grammatical sketch. In Jacobs, J., Stechow, A., Sternefeld, W., and Vennemann, T., editors, *Syntax: An International Handbook of Contemporary Research*, volume 9.2 of *Handbooks of Linguistics and Communication Science*, pages 1288–1319. de Gruyter.

Gurevych, I. and Niederlich, H. (2005). Computing semantic relatedness in German with revised information content metrics. In *Proceedings of OntoLex 2005 - Ontologies and Lexical Resources IJCNLP 05 Workshop*, pages 28–33, Jeju Island, Republic of Korea.

Harras, G. (1989). Zu einer Theorie des lexikographischen Beispiels. In Hausmann, F.J. et al., editor, *Dictionaries: An International Encyclopedia of Lexicography*, Handbooks of Linguistics and Communication Science, 5.1, pages 607–614. de Gruyter.

Hartmann, R. (2001). *Teaching and Researching Lexicography*. Longman, Harlow.

Hawkins, J. (1992). Syntactic weight versus information structure in word order variation. In Jacobs, J., editor, *Informationsstruktur und Grammatik*. Westdeutscher Verlag, Opladen.

Hawkins, J. (1994). *A Performance Theory Of Order And Constituency*. Cambridge University Press, Cambridge.

Herbst, T. (1989). Dictionaries for foreign language teaching: English. In Hausmann, F.J. et al., editor, *Dictionaries: An International Encyclopedia of Lexicography*, Handbooks of Linguistics and Communication Science, 5.2, pages 1379–1385. de Gruyter.

Hermanns, F. (1988). Das lexikographische Beispiel. Ein Beitrag zu seiner Theorie. In Harras, G., editor, *Das Wörterbuch. Artikel und Verweisstrukturen*, Sprache der Gegenwart. Band LXXIV, pages 161–195. Schwann.

Higgins, D. (2005). Which statistics reflect semantics? Rethinking synonymy and word similarity. In Kepser, S. and Reis, M., editors, *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, Studies in Generative Grammar 85, pages 265–284. Mouton De Gruyter, Berlin.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press, Cambridge, MA.

Howell, D. C. (2002). *Statistical Methods for Psychology*. Duxbury, Pacific Grove, CA, 5th edition.

Hunt, K. (1965). Grammatical structures written at three grade levels. Research Report No. 3.

Jackson, H. (2002). *Lexicography: An Introduction*. Routledge, London.

Jacobsen, J., Manley, J., and Pedersen, V. (1991). Examples in the bilingual dictionary. In Hausmann, F.J. et al., editor, *Dictionaries: An International Encyclopedia of Lexicography*, Handbooks of Linguistics and Communication Science, 5.3, pages 2782–2789. de Gruyter.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 212–219, Borovets, Bulgaria.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33, Taipei, Taiwan.

Kanejiya, D., Kumar, A., and Prasad, S. (2003). Automatic evaluation of students' answers using syntactically enhanced LSA. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, pages 53–60, Edmonton, Canada.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29:3:459–484.

Kempcke, G. (1992). Organisationsprinzipien und Informationsangebote in einem Lernerwörterbuch. In Brausse, U. and Viehweger, D., editors, *Lexikontheorie*

und Wörterbuch, Lexicographica: Series maior; 44, pages 165–244. Niemeyer, Tübingen.

Landauer, T. (2002). On the computational basis of learning and cognition: Arguments from LSA. In Ross, N., editor, *The Psychology of Learning and Motivation*, pages 43–84. Academic Press, San Diego.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Langenscheidt (1991). *Basic German Vocabulary*. Langenscheidt.

Langenscheidt (2003). *Großwörterbuch Deutsch als Fremdsprache*. Langenscheidt, Berlin.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1085–1090, Edinburgh, Scotland.

Laufer, B. (1992). Corpus-based versus lexicographer examples in comprehension and production of new words. *EURALEX '92 - Proceedings*, pages 71–76.

Lenz, A. (1998). *Untersuchungen zur Beispiel- und Beleglexikographie historischer Bedeutungswörterbücher unter besonderer Berücksichtigung der Neubearbeitung des Deutschen Wörterbuchs gegründet von Jacob und Wilhelm Grimm*. PhD thesis, Georg-August-Universität, Göttingen. Available online: http://webdoc.gwdg.de/diss/2001/lenz/diss.pdf [2002, Feb 5].

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine from an ice cream cone. In Gleitman, L. and Josh, A., editors, *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Ontario.

Li, J., Bandar, Z., and McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15 (4):871–881.

Li, J., Bandar, Z., McLean, D., and O'Shea, J. (2004). A method for measuring sentence similarity and its application to conversational agents. In Barr, V. and Markov, Z., editors, *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, pages 820–825, Miami Beach, FL. AAAI Press.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28 (2):203–208.

Maingay, S. and Rundell, M. (1987). Anticipating learner's errors - implications for dictionary writers. In Cowie, A., editor, *The Dictionary and the Language Learner*, Lexicographica: Series maior ; 17, pages 128–135. Niemeyer, Tübingen.

Maingay, S. and Rundell, M. (1990). What makes a good dictionary example? Paper presented at the 24th IATEFL Conference, Dublin.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marello, C. (1987). Examples in contemporary Italian bilingual dictionaries. In *The Dictionary and the Language Learner*, Lexicographica: Series maior ; 17, pages 224–237. Niemeyer, Tübingen.

Martin, R. (1989). L'exemple lexicographique dans le dictionnaire monolingue. In Hausmann, F.J. et al., editor, *Dictionaries: An International Encyclopedia of Lexicography*, Handbooks of Linguistics and Communication Science, 5.2, pages 599–607. de Gruyter.

McDonald, S. (1997). A context-based model of semantic similarity. Unpublished.

Menard, S. (1995). *Applied Logistic Regression Analysis*. Number 07-106 in Sage university paper series on quantitative applications in the social sciences. Sage Publications, Thousand Oaks, CA.

Miller, G. and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6 (1):1–28.

Miller, G. and Chomsky, N. (1963). Finitary models of language users. In Luce, R., Bush, R., and Galanter, E., editors, *Handbook of Mathematical Psychology, Vol.2*, pages 419–491. Wiley, New York.

Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition*, 25:473–499.

Mondria, J.-A. and Wit-De Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, 12:249–267.

Morris, J. and Hirst, G. (2004). Non-classical lexical semantic relations. In *Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the ACL*, Boston.

Moulin, A. (1983). The pedagogical/learner's dictionary: The LSP learner's lexicographical needs. In Hartmann, R., editor, *Lexicography: Principles and Practice*, pages 144–152. Academic Press, London.

Nesi, H. (1996). The role of illustrative examples in productive dictionary use. *Dictionaries: The Journal of the Dictionary Society of North America*, 17:198–206.

Neubauer, F. (1989). Vocabulary control in the definitions and examples of monolingual dictionaries. In Hausmann, F.J. et al., editor, *Dictionaries: An International Encyclopedia of Lexicography*, Handbooks of Linguistics and Communication Science, 5.1, pages 899–905. de Gruyter.

Nikula, H. (1986). Wörterbuch und Kontext. Ein Beitrag zur Theorie des lexikalischen Beispiels. In Schöne, A., editor, *Kontroversen, alte und neue: Akten des VII. Internat. Germanisten-Kongresses, Göttingen 1985*. Niemeyer, Tübingen.

Olde, B. A., Franceschetti, D. R., Karnavat, A., and Graesser, A. C. (2002). The right stuff: Do you need to sanitize your corpus when using Latent Semantic Analysis? In *Proceedings of 24th Annual Meeting of the Cognitive Science Society*, pages 708–713, Hillsdale. Erlbaum.

Padó, S. (2002). Extracting semantic information from corpora using dependency relations. MSc Thesis.

Padó, S. and Lapata, M. (2003). Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Comutational Linguistics (ACL)*, pages 311–318, Philadelphia.

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. John Benjamins, Amsterdam.

Parker, K. and Chaudron, C. (1987). The effects of linguistic simplification and elaborative modifications on L2 comprehension. Paper presented at the 21st Annual TESOL Convention.

PONS (2004). *PONS Großwörterbuch Deutsch als Fremdsprache*. Klett, Stuttgart, 3rd edition.

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press, Cambridge.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., and Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25:337–354.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference of Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, Canada.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal Artificial Intelligence Research*, 11:95–130.

Rubenstein, H. and Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8 (10):627–633.

Sahlgren, M. (2001). Vector-based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*, page 1036, Helsinki, Finland.

Schouten-van Parreren, C. (1989). Vocabulary learning through reading: Which conditions should be met when presenting words in texts? *AILA Review (Vocabulary Acquisition)*, 6:75–85.

Serafin, R. and Di Eugenio, B. (2004). FLSA: Extending Latent Semantic Analysis with features for dialogue act classification. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 692–699, Barcelona, Spain.

Shimohata, M. (2004). *Acquiring Paraphrases from Corpora and its Application to Machine Translation*. PhD thesis, Nara Institute of Science and Technology.

Sinclair, J. (1987). Introduction. In Sinclair, J., Hanks, P., Fox, G., Moon, R., and Stock, P., editors, *Collins Cobuild English Language Dictionary (First edition)*. Collins, London and Glasgow.

Smith, C. (1988). Factors of linguistic complexity and performance. In Davison, A. and Green, G., editors, *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, pages 247–279. Lawrence Erlbaum, Hillsdale, N.J.

Sotillo, S. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning & Technology*, 4 (1):82–119.

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502, Freiburg, Germany.

UIS-Parser (2002). UIS-parser. Available online at: http://www.ifi.unizh.ch/CL/UIS/parser.html [2003, Dec 16].

Ullman, M. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition*, 4:105–122.

Uszkoreit, H. (1987). *Word Order And Constituent Structure In German*. CSLI, Stanford.

Van Parreren, C. and Schouten-Van Parreren, M. (1981). Contextual guessing: A trainable reader strategy. *System*, 9,3:235–241.

Vavra, E. (2000). Definitions of the T-unit. Pennsylvania College of Technology. Retrieved [2003, Jan 14] from the World Wide Web: http://nweb.pct.edu/homepage/staff/evavra/ED498/Essay009_Def_TUnit.htm.

WAHRIG (2003). *WAHRIG Deutsches Wörterbuch*. Wissens Media Verlag GmbH, Gütersloh, CD-ROM edition.

Wang, M. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, 9:398–404.

Warschauer, M. (1996). Comparing face-to-face and electronic communication in the second language classroom. *CALICO*, 13:7–26.

Wiemer-Hastings, P. (1999). How latent is Latent Semantic Analysis? In *Proceedings of the 16th International Joint Congress on Artificial Intelligence*, pages 932–937, San Francisco. Morgan Kaufmann.

Wiemer-Hastings, P. (2000). Adding syntactic information to LSA. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 989–993, Mahwah, NJ. Erlbaum.

Wiemer-Hastings, P. (2004). All parts are not created equal: SIAM-LSA. In *Proceedings of 26th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum.

Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In *Artificial Intelligence in Education*, pages 535–542. IOS Press, Amsterdam.

Wiemer-Hastings, P. and Zipitria, I. (2001). Rules for syntax, vectors for semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 1112–1117, Mahwah, NJ. Erlbaum.

Xu, H. (2005). Treatment of deictic expressions in example sentences in English learner dictionaries. *International Journal of Lexicography*, 18(3):289–311.

Yngve, V. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, 10,3:85–101.

Zgusta, L. (1971). *Manual of Lexicography*. Mouton, The Hague.

Zipf, G. (1949). *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA.

Zipitria, I., Elorriaga, J. A., and Arruarte, A. (2006). LSA learner sentence comprehension in agglutinative and non-agglutinative languages. Workshop on Teaching With Robots, Agents, and NLP. ITS 2006. Available: http://facweb.cs.depaul.edu/elulis/zipitria.pdf [2006, Aug 6].

Zöfgen, E. (1986). Kollokation, Kontextualisierung, (Beleg-)Satz. Anmerkungen zur Theorie und Praxis des lexikographischen Beispiels. In Barrera-Vidal, A. et al., editor, *Französische Sprachlehre und bon usage*, pages 219–238. Hueber, München.

Zöfgen, E. (1994). *Lernerwörterbücher in Theorie und Praxis: ein Beitrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen*. Lexicographica. Series maior, 59, Tübingen.