

Predicting linking elements in German compounds: A stochastic approach

Torsten Marek shlomme@gmx.net

June 2005

Abstract

German compounds usually take a semantically empty linking element (or morpheme) between head and complement (Fugenelement, like the *-s-* in *Mittag-s-zeit*). Some analyses about the choice of different linking elements (like *-en-*, *-s-*, *-e-*) in compounds are based on semantic features or inflexion classes of head and complement. Such knowledge is generally not available when the prediction is to be made by a computer program, an easier theory is desirable.

[Baa03] showed a probabilistic approach on predicting linking elements relying only on phonological features of head and complements for Dutch, which produces quite useful results. This paper tries to apply Baayen's theory on German compounding. The new formalism's performance will be evaluated to find out if it is similar to the performance observed for Dutch.

Contents

1	Approaches in the literature	2
1.1	German: Combining phonology, morphology and the lexicon . . .	2
1.2	Dutch: Guessing from phonological features	3
2	Motivation for the stochastic approach	5
3	Testing the stochastic approaches	6
3.1	Gathering example data	6
3.2	Spreading activation approach	7
3.3	The IB1-IG distance measure	8
3.4	Feature selection	8
3.5	Discussion of the results	8

1 Approaches in the literature

1.1 German: Combining phonology, morphology and the lexicon

[Fuh96] gives a concise view over German linking elements with regard to phonological, morphological and lexical motivation for all observed linking elements. To be able to see whether a linking element actually occurs in a compound, the author gives defines which word form of a lexical category is used as a base for composition in German.

verbs: the verbal root. Example: *laufen* → *lauf-*

nouns: the nominative singular form

adjectives: the “simple form” (uninflected)

In this paper, we are only interested in nouns. In nouns compounds, the following linking morphemes occur: *-s-*, *-es-*, *-ə-*, *-n-*, *-en-*, *-er-* or *-∅-*. The initial problem is that the choice of a linking morphemes for a given complement-head combination is not predictable by complement or head only, as some examples show: *Kind-s-kopf* but *Kind-er-garten*, *Kranke-n-kasse* but *Spar-kasse* (One important remark: the dashes in the words separate the linking morphemes from head and compound, they are not necessarily syllable boundaries!). The process for selecting the linker therefore has to be more complex and not only lexically defined, since linking morphemes are productive and speakers of German have an intuitive notion of what the right linking morpheme when when they are confronted with the task of composing novel words.

The author describes the behaviour and constraints of linking elements with respect to the different layer of word formation ([Fuh96, pp. 527]).

Phonology

There seems to be some evidence for prosodic motivation of linking elements in certain contexts, where they prohibit adjacent stressed syllables. Another observation is that all linking morphemes always close a syllable, due to their high sonority they cannot be taken into the onset of the head word.

Morphology

The triggered linking morphemes often mimic the nominative plural or genitive singular form of the complement noun, if only diachronically (*Schwan-en-hals* (today the genitive is *Schwan(e)s*). The Author makes an important distinction between paradigmatic (i.e. linking morphemes “looking like” a form of the noun paradigm) and unparadigmatic linking elements, where *-s-* is the only one appearing unparadigmatically ([Fuh96, p. 529]). If the linking morpheme is only diachronically paradigmatic, the assumption is that the compound is lexicalized (*Schmerz-ens-geld* being another example).

Lexical rules

In the discussion of the different linking elements a lot of lexical triggers for specific linking elements are mentioned. If the complement is a suffixed feminine noun (*die Kindheit, die Versicherung*) it takes the (unparadigmatic) *-s* (*Kindheit-s-trauma, Versicherung-s-mathematiker*), and so on. Listing those rules is tiresome and, as we will see later, does not meet our specifications for a usable prediction mechanism.

Conclusion

The author's conclusion is that although linking elements do not appear randomly, the formulation of proper rules is not easy, and sometimes impossible. However, there is hope on the way, since new compounds are created with default linking morphemes instead of paradigmatic ones, which makes predicting them without intense background knowledge a lot easier.

1.2 Dutch: Guessing from phonological features

The approach of [Baa03] to linking elements is different: the scope of the article (fragment) is limited to predicting linking elements in noun-noun compounds from phonological features of head and complement. The problems in Dutch compounding are roughly the same as in German, the choice of linking elements is not determined by head or complement only; they are productive as well. Whether linking morphemes can be morphemes from the noun paradigm or whether they are triggered by other lexical features is not discussed at all. The author also states that earlier research did not produce an near-exhaustive rule system that can describe all phenomena correctly. He therefore proposes an alternative approach: instead of consuming all examples and creating a pure and abstract cascade of rules (which would also have to allow a lot of exceptions, as we already saw for German), predicting linking morphemes for a novel compound should be done by looking at all example compounds and selecting the pattern that matches best. The further approach is called syntagmatic with greedy learning (= examples are consumed and forgotten), the latter one paradigmatic with lazy learning (= examples are stored for later usage).

For the prediction, five distinctive features of a noun-noun combination are selected:

- the complement word
- the head word
- the nucleus of the complement word (a vowel or diphthong)
- the onset of the head word
- the coda of the head word

[Baa03, Table 7.1]

The combination of all these five features is associated with exactly one linking element.

Given a new combination of words, the most probable linking element is that, whose associated pattern does not or only insignificantly differ from the feature set of the new combination. The distance measure is the simple “Hamming distance”, that is: Given two equally long, arbitrary strings X, Y , the Hamming distance $\Delta(X, Y)$ is the number of different elements at the same position. Or, the formal definition ([Baa03, p. 246]):

$$\Delta(X, Y) = \sum_{i=1}^n I_{[x_i \neq y_i]}$$

The values of the X ’s features are denoted by x_i (similar for Y), and the function $I_{[z]}$ maps the truth values $\{True, False\}$ of a boolean expression to $\{1, 0\}$. Since not all features have the same influence on the selection of a linking morpheme, the author motivates the usage of a weight for each feature. The weight is based on the entropy of this element in the set of examples. (I will leave out the discussion of entropy here, we will return there when the weighing of the features for German is discussed). The weighed version of the formula is:

$$\Delta(X, Y) = \sum_{i=1}^n w_i I_{[x_i \neq y_i]}$$

where w_i is the relative weight of the feature. Tweaking of the weights shows that modifier and head have the highest influence on the selection of a linking element. In contrast to that, other features like stress on the first constituent have no meaningful information ([Baa03, p. 248]). These findings have been supported by experimental studies. The algorithm is claimed to predict around 92% of all linking elements in the author’s test correctly. This algorithm is also called the IB1-IG distance measurement.

The other algorithm described by the author in [Baa03] and [KSB02] is the so-called spreading activation model. In contrast to the upper approach, a similarity rather than a distance measurement is taken, but the algorithms are theoretically equivalent. Given a compound with a complement c , a head h and an unknown linking morpheme l the spreading activation algorithm just scans the whole example database for words with head h or complement c . It creates a histogram of all occurring linking morphemes in these words, and the most probable linking morpheme is chosen. Occurrences of linking morphemes can be weighed with respect to the word being matched in head or complement and the overall frequency of the word. The former one is simple, but the latter one is computationally more intensive and, above all, needs reliable frequency values for each word, therefore it is not implemented.

2 Motivation for the stochastic approach

Both approaches provide a usable formalism to predict the distribution of linking morphemes in simple German noun-noun compounds. A choice between the two has to be made and I will briefly outline what kind of data structures and algorithms will be needed. It will clarify that the choice will be the stochastic approach, since there are good reasons from a computational point of view. To implement the rule-driven part of [Fuh96] one would need:

- an exhaustive lexicon of German, with information about inflexion paradigm, gender and stress
- morphological analyzers for derived words
- a rule system to combine the latter

The simple example for *-s-* as the linker when the complement is a derived female noun:

$$is_derived(complement) \wedge is_female(complement) \Rightarrow -s-$$

Where *is_derived* knows about all derivative suffixes (like *-heit*, *-ung* etc.) and *is_female* combines lexicon and morphological analysis to check the gender. The morphological analysis could be spared out, but this only leads to more lexicon lookups and a larger lexicon.

In contrast, an implementation of [Baa03]

- an example database with splitted compounds
- feature weights
- an program that combines database lookup and the stochastic algorithm

The spreading activation algorithm is even simpler, because the splitted compounds suffice and no further features are needed. If the word frequencies are left out, the spreading activation model is also much faster. The distribution of linking morphemes can be computed and cached for every feature once and forever. For a given set of features the histograms have to be added up. Thus, this algorithm has complexity $O(n)$ (where n is the number of feature sets looked up). In the other algorithm, a feature set has to be compared to every other feature set in the sample data base, resulting in $O(mn)$ (where m is the number of samples).

Both approaches are data-centric (but use different data), so comparing the amount of storage needed does not help (and many linguists and computational will wrongly argue that storage and speed of access does not matter any more). The advantage of the stochastic approach is its simplicity: all the knowledge resides in the example database and the weights. The example database can easily be created by programs and if the feature set changes, the features can be extracted from the words by (the well-understood) finite state automatons

using one of the very advanced finite state toolkits or regular expressions as seen in many programming languages, notably Perl. The weights can be estimated (which we will do later) or calculated using supervised or unsupervised machine learning, given a large and representative example base. Therefore, changing the feature just triggers a new run of the learning algorithm but does not impose programmers (or linguists) with recalculating the weights by hand. With the rule-based system, there is not only a lexicon that has to be maintained (there are enough digital lexicons for German), but also the system of rules needs to be written and updated with new knowledge, leading to more code compared to the stochastic implementation.

3 Testing the stochastic approaches

3.1 Gathering example data

To be able to test an algorithm, we need a considerable amount of correctly splitted German noun-noun compounds. As a base, I used the German wordlist for Ispell and Aspell ¹, with approx. 80,000 base words. A little script was written to extract compound words from the list and split them correctly into complement, linking morpheme and head. Depending on the strategy and the source lists, the output varied heavily in size and quality, and some small corrections were made to the word list to improve the output. The overall distribution of linking morphemes nonetheless showed to be quite stable:

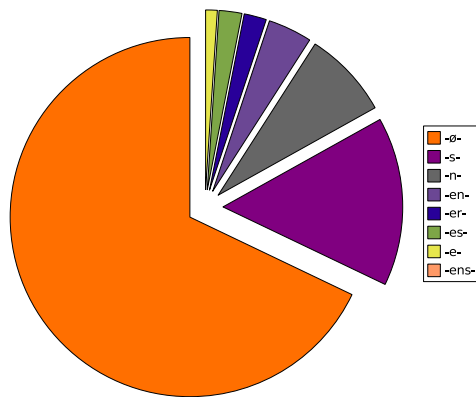


Figure 1: Morpheme distribution

morpheme	frequency
-∅-	68%
-s-	15%
-n-	8%
-en-	4%
-er-	2%
-es-	2%
-e-	1%
-ens-	0.05 %

Table 1: Distribution table

Unfortunately I was not able to find reliable (i. e. with a comparably large data base) sources to compare the number. The only actual number I read is

¹<http://j3e.de/ispell/igerman98/>

that 35%². The sum of compounds with nonzero linking elements is around 32%. The gained data is, of course, noisy. This is partly the fault of the word list (quite some base word forms are missing), and partly the fault of the split guesser. Every compound that begins with *Eisen* will be split into *Eis-en-*, since the algorithm prefers long linking elements over short ones. This is by far not the only source of error. Rough measurements show that the error might be around 5% (which is acceptable), and every algorithm has to be able to cope with partially wrong data anyhow.

3.2 Spreading activation approach

Before caring about the selection of features a simple script was used to predict the linking elements based on the spreading activation approach. Only head and complement were used as features of the compound but even this gives quite good results: The 20,000 example compounds were randomly split into

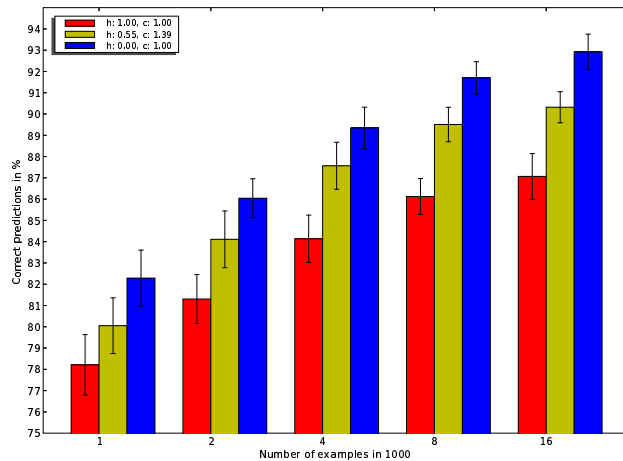


Figure 2: Prediction performance for spreading activation approach

1,000 test compounds and training compounds. The x axis shows the size of the sample data base. For each sample data base size, three different combinations of weighing factors were taken (given in the legend). Every combination (weighing factors, sample data base size) was run 50 times. The diagram shows the mean percent of correctly predicted linking morphemes, together with the standard deviation. When the features are weighed according to the formula in [KSB02, p. 716], the complement gets a higher weight as the head (which is in line with our intentions). In the third setting, the head has the weight zero. The linking

²<http://jocn.mitpress.org/cgi/content/full/16/9/1647>, although the number does not appear in the quoted literature

morpheme is chosen solely by the complement. Surprisingly, the performance is slightly better. This might hint that instead of the full head, only the onset should be chosen as a feature with high weight.

3.3 The IB1-IG distance measure

A simple implementation of the algorithm using the Hamming distance measurement also does not take much time, yet initially runs much slower since a given compound with unknown linking element has to be compared to every other feature set in the sample data base. Instead of writing an own version of this algorithm the free program TiMBL³ by Daelemans et al. was used. This program is highly optimized for stochastic distance measurements and has been used in [Baa03] as well.

3.4 Feature selection

Possible features for the selection of a linking morpheme are:

- the complement word (already shown to be distinctive)
- the head word (already shown to be distinctive)
- the nucleus of the complement
- the coda of the complement
- the onset of the head

For different combinations of features a TiBL test was run. Two lists of splitted compounds were taken from the example data base and annotated with the needed additional features. One was used as training data, the other as test harness. The training list usually was larger than the test list. As it is possible to see, the accuracy does not depend on the actual ratio between the length of the two lists but on the size of the training data.

The legend contains the experiment setup: the features (C is complement, H is head) and the sample data base size (last number). For each setup, 4 experiments were made, and the mean result was taken.

3.5 Discussion of the results

As the accuracy numbers for the selected feature sets show there is no magic feature that drives the prediction precision upwards. The best feature combination is (complement, head, coda(complement)). But even with them, as further test show, around 4000 examples are needed to get a prediction precision of around 90%. Still, the example data could be better, and more feature sets can

³Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch (2004). TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report 04-02, Available from <http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf>

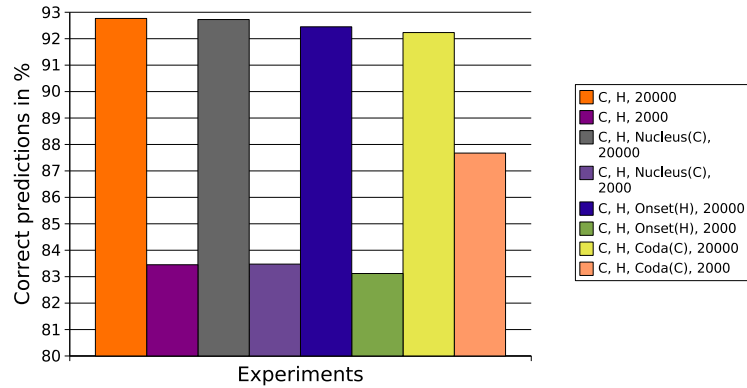


Figure 3: Prediction accuracies for different feature configurations

be explored. But we need to keep in mind that there will always be a number of compounds we can only predict with deep lexical knowledge. One example is *Herz-ens-güte*, which is definitely a lexicalized compound, since the linking element *-ens-* is long gone from the inflexion paradigm for *Herz*. Treating this and other compounds as lexicalized is feasible, since three most common linking morphemes (*-Ø-*, *-s-* and *-n-*) together make up around 90% of all linking morphemes used in German. Using that knowledge a program might be able to predict linking morphemes for nearly all German compounds correctly.

References

- [Baa03] R. Harald Baayen. Probabilistic approaches to morphology. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic linguistics*. MIT Press, 2003.
- [Fuh96] Nanna Fuhrhop. Fugenelemente. In Ewald Lang and Gisela Zifonun, editors, *Deutsch - typologisch*. de Gruyter, 1996.
- [KSB02] Andrea Krott, Robert Schreuder, and R. Harald Baayen. Linking elements in dutch compounds. In Harry A. Whitaker, editor, *Brain and language*, volume 81. Academic press, 2002.