Published in: Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology, Sep 30–Oct 2 2015, University of Duisburg-Essen, Germany - Gesellschaft für Sprachtechnologie and Computerlinguistik e. V., Pp. 49–57.

Growing trees from morphs: Towards data-driven morphological parsing

Petra Steiner and Josef Ruppenhofer
Institute of Information Science and Natural Language Processing
Hildesheim University
Hildesheim, Germany
{ruppenho,steinerp}@uni-hildesheim.de

Abstract

We present a quantitative approach to disambiguating flat morphological analyses and producing more deeply structured analyses. Based on existing morphological segmentations, possible combinations of resulting word trees for the next level are filtered first by criteria of linguistic plausibility and then by weighting procedures based on the geometric mean.

The frequencies for weighting are derived from three different sources (counts of morphs in a lexicon, counts of largest constituents in a lexicon, counts of token frequencies in a corpus) and can be used either to find the best analysis on the level of morphs or on the next higher constituent level. The evaluation shows that for this task corpus-based frequency counts are slightly superior to counts of lexical data.

1 Introduction

One of the bottlenecks for the automatic processing of German language data is word form productivity. For the specification of concepts, the creation of long compounds and derived forms is very common, e.g. (1).

(1) Oberklassenschlagbohrmaschine
'Premium class hammer drill (machine)'

While constituents of English compounds are often separated by hyphens or spaces, in German the constituents of compounds are written as a single orthographic word. Thus, the word form in (1) could be analyzed (usefully) as *Oberklasse* 'premium class', and *Schlagbohrmaschine* 'hammer drill', but also (uselessly) as *Ober* 'premium', *Klassenschlag* '*class hit', *bohr* 'drill', and *Maschine* 'machine'. Note the interfix *n* between *klasse* and *schlag*. Furthermore, some morphs can be ambiguous. E.g. *Ober* might denote a waiter, while *Schlag* might be related to the verb *schlagen* 'hit, hammer' or the noun *Schlag* 'hit, blow'. Moreover, the spelling conventions of German result in ambiguity concerning morph boundaries. E.g., *Anbaumenge* could be analyzed into the immediate constituents *Anbau* 'cultivation' and *Menge* 'amount' but also to *An* 'at', *Baum* 'tree' and *Enge* 'narrowness'.

Applications in machine translation or multilingual terminology extraction require robust methods for disambiguating and post-processing morphological analyses of German words. While some robust morphological analyzers for German exist (e.g. SMOR (Schmid et al., 2004), Gertwol (Haapalainen and Majorin, 1995), MORPH (Hanrieder, 1991; Hanrieder, 1996), TAGH (Geyken and Hanneforth, 2006)), all of them yield flat structures. However, hierarchical word structures provide important information about a word's meaning and should be taken into account as well.

Some heuristics, such as taking the analysis with the smallest number of constituents, can be used to inform the choice between multiple analyses. However, there is room for refinement and augmentation. Cap (2014) discusses a broad range of approaches to disambiguating compounds. The present contribution, in contrast, aims at dealing not only with compounding but also with derivation and other word formation processes.

Würzner and Hanneforth (2013) tackle the problem of full morphological parsing, but restricted to adjectives. They segment words into lexical units using the TAGH system of (Geyken and Hanneforth, 2006) and then use a probabilistic context free grammar for parsing. The grammar is trained on manually labeled word trees. Our approach is more general in that we cover complex words of any part of speech. It is more limited in that we do not produce a full parse.

Most importantly, since the morphology system we use, SMOR, produces more segmentations per item than TAGH, we focus on disambiguation of available analyses, for which we also use corpus frequency counts, unlike Würzner and Hanneforth (2013).

Our approach starts from sets of flat analyses, builds all possible combinations of higher-level analyses, and filters these using the geometric mean (*gm*) score. In the calculation of the score, we use either frequencies derived from lexicons or frequencies derived from corpora:

a) all morphs found in a German lexicon with their frequencies within the lexicon,

b) all immediate constituents found in a German lexicon with their frequencies derived from the lexicon,

c) all immediate constituents found in a German lexicon with their frequencies taken from a German corpus.

Section 2 presents the data and their pre-processing and augmentation. Section 3 describes our gold standard. The methods for weighting and filtering the morphological analyses are presented in Section 4, which also shows our approach to handling data sparsity. The results for each of the three datasets are presented in Section 5 and discussed in Section 6. The last section comprises a conclusion with an outlook for future work.

2 Data

2.1 Augmented SMOR Analyses

SMOR is a morphological analyzer based on two-level morphology (Koskenniemi, 1984), implemented as a set of finite-state transducers (Schmid et al., 2004). For German, a large set of lexicons is available. The final version used for the current work comprises a main lexicon with 41,944 entries, proper name lexicons with 15,188 entries and different datasets with other morphological information. These lexicons contain information about inflection, parts of speech and classes of word formation (e.g. abbreviations, truncations). The tag set used is compatible with the STTS (Stuttgart Tübingen tag set, Schiller et al. (1995)).

The output for (1) with information on word formation and inflection is given in Figure 1. Please note that the interfix between *Klasse* and *schlag* has been deleted in these analyses by SMOR. Also, the STTS-like annotation contains some metatags for abbreviations and word-form parts before or between hyphenation as in example (2), which is a hyphenated variant of (1).

(2) {Oberklassen}-<TRUNC>Schlag<NN>bohren <V>Maschine<+NN><Fem><Acc><Sg>

This leaves part of the word unanalyzed and is an unwanted side effect. We therefore reanalyze results with tag <TRUNC> as follows:

a) Hyphens are removed, the letters following the hyphens are transformed to lower case. The copy is used as input of SMOR. If an analysis was found, hyphens and letters are re-inserted.

b) If only an analysis with <TRUNC> is possible, each string between hyphens is reanalyzed separately. For this process, the SMOR lexicons are reused.

This leads to analyses such as (3).

(3) Ober<PREF>Klasse<NN>n<FL>-<HYPHEN> Schlag<NN>bohren<V>Maschine<+NN>

Interfixes are restored from internal SMOR results[1] and annotated with *FL* (filler letter) as a new tag for the interfix. Table 1 summarizes the changes.

Method	*t*	*n*	*a*	*r*
(a) SMOR baseline	105	3	0	0.00
(b) remove hyphens	48	2	58	0.54
(c) reanalyze TRUNC	39	3	66	0.61
(d) combine (b) and (c)	2	2	104	0.96

Table 1: Analyzed hyphenated forms; *t* analyses containing TRUNC; *n*: hyphenated forms without analyses; *a*: correctly pre-analyzed hyphenated word form; *r*: relative frequency of *a*

We used the 1,101 items from our gold standard data (see Section 3). Of the 108 word forms containing one or more hyphens, only two were not covered by any method. The methods (b) and (c) work rather complementarily. The analyses of hyphen-removed forms are especially successful for spelling variants, such as *anti-amerikanisch* ‘anti-American’. On the other hand, the lexicon-based analyses cover especially word forms which

[1] Sennrich and Kunz (2014) also add interfixes to SMOR output.

```
ober<PREF>Klasse<NN>Schlag<NN>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN>Schlag<NN>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN>Schlag<NN>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN>Schlag<NN>bohren<V>Maschine<+NN><Fem><Nom><Sg>
ober<PREF>Klasse<NN>schlagen<V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN>schlagen<V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN>schlagen<V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN>schlagen<V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Nom><Sg>
```

Figure 1: Output of SMOR for *Oberklassenschlagbohrmaschine*

include abbreviations, e.g. *CO2-Emissionen* 'CO2 emissions'. When hyphenated word forms which cannot be analyzed after removal of the hyphens are processed by the second algorithm, only four hyphenated word forms remain unanalyzed, due to misspelling or unusual forms that were not included in the SMOR lexicon.

Small changes to the lexicon, such as adding proper names or changing restrictions on morph positions inside words, allow for complete coverage of the observed data. The analyses are reduced to the lemma form. The sequence of the morphological information is transformed by using directed acyclic graphs, resulting in output such as (4), giving a surface form and a lexical form of the word analysis, followed by the tags.

```
(4)    Ober  klasse  n   schlag  bohr    maschine
       ober  Klasse  n   Schlag  bohren  Maschine
       PREF  NN      FL  NN      V       NN          <NN>
```

2.2 CELEX

The lexical database CELEX contains Dutch, English, and German lexical information (Baayen et al., 1995) combined with frequency information, which for German is based on counts of the *Mannheim Corpus* (Gulikers et al., 1995, 102ff.). The morphological part is of special interest for word analyses (Gulikers et al., 1995, 45ff.). The database gives information on word-formation types and provides manually annotated multi-level word structures from which flat as well as complex structures can be extracted. Special characters of German such as *ä* and *ß* are represented as *"a* and *$* in the lexical part of CELEX and had to be changed. Information about orthography is available in the database. However, it is restricted to lemmas. Therefore, the components of morphological analyses had to be adapted heuristically and were manually corrected. All ablauts which occur in irregular verbs were changed manually.

In total, our modified CELEX dataset for German has 51,727 entries. From it, three datasets with frequency information were extracted:

a) all morphs with their frequencies within the CELEX lemmas,

b) all immediate constituents with their frequencies within the CELEX lemmas,

c) all immediate constituents within the CELEX lemmas with their frequencies as found in the *Mannheim Corpus*.

For example, the lemma *Sprachwissenschaft* increments the frequencies for each of the morphs *sprech* (*Sprache* is a derivative of *sprechen*), *wissen* and *schaft* by 1. Likewise, the frequencies of its immediate constituents, *Sprache* and *Wissenschaft*, are incremented by 1. For the dataset of the text frequencies, 13 is added for each of the immediate constituents, as this is the lemma's corpus frequency. This leads to 13,419 entries for the morphs and their frequencies, and 21,406 entries for the constituents and their frequencies within the lexicon and the corpus.

The first dataset is used to choose among the best morph-level analyses, the other frequency data provide input for higher-level analyses.

3 Gold Standard

The gold standard used is based on Cap (2014, 95), who uses part of the test set of the 2009 workshop on statistical machine translation.[2] Of these 6,187 tokens, 1,101 were analyzed by human annotators, as in (5).

(5) 10-Jahres-Prognosen 10|Jahr|Prognosen
'10-year forecast'

These compounds are input for the analyses of morphological structure. The analysis of the lemmatized form with hyphens and interfixes in Cap (2014) included forms like (6), which made it necessary to create a new gold standard for our evaluation (cf. Section 5).

[2] http://www.statmt.org/wmt09/translation-task.html

(6) 10|-|Jahr|es-|Prognose

4 Methods

4.1 Word structures as Integer Compositions

The combinatorial structure of morphological analyses of a word with *n* parts is isomorphic to the permuted integer partitions of *n*. For instance, a word which is analyzed into three noun stems can be described in four different ways (7a–d). While (7a) shows an analysis of a syntagmatic compound (German: Zusammenrückung), in (7b) and (7c) the immediate constituents *Drahtseil* and *Seilakt* are identified. (7d) interprets the three-stem analysis as incorrect and amalgamates them to a monomorphemic word. The correct analysis for immediate constituents is (7c), for the smallest units (morphs) it is (7a).

(7) *Drahtseilakt* ‘High-wire act’

a. [[’Draht’], [’seil’], [’akt’]]
b. [[’Draht’], [’seilakt’]]
c. [[’Drahtseil’], [’akt’]]
d. [[’Drahtseilakt’]]

The isomorphic structure of integer compositions shows the number of elements in the subsets of the sequential elements of each morphological analysis (cf. (8)). The algorithm for processing the combinatorially possible analyses makes use of this analogy.

(8) Integer compositions corresponding to the analyses in (7) above

a. 1-1-1
b. 1-2
c. 2-1
d. 3

The number of all integer compositions for *n* equals 2^{n-1} for integers $>= 1$. For *Oberklassenschlagbohrmaschine* with *n* = 5 this gives 16 compositions. The interfix does not count as a relevant morph.

However, some compositions can be ruled out as linguistically implausible, e.g. compositions starting with a suffix or ending with a prefix. This does not only reduce the number of combinatorially possible analyses but also splits the set into subsets marked by affix boundaries. E.g. some compositions for *abwechslungsreich* ‘rich in variety’ yield impossible subcomponents such as **ungsreich* ‘SUFFIX FL full’ in (9b) and (9f).

(9) Compositions of *abwechslungsreich*

a. [[’ab’], [’wechsl’], [’ung’, ’s’], [’reich’]],
b. [[’ab’], [’wechsl’], [’ung’, ’s’, ’reich’]],
c. [[’ab’], [’wechsl’, ’ung’, ’s’], [’reich’]],
d. [[’ab’], [’wechsl’, ’ung’, ’s’, ’reich’]],
e. [[’ab’, ’wechsl’], [’ung’, ’s’], [’reich’]],
f. [[’ab’, ’wechsl’], [’ung’, ’s’, ’reich’]],
g. [[’ab’, ’wechsl’, ’ung’, ’s’], [’reich’]],
h. [[’ab’, ’wechsl’, ’ung’, ’s’, ’reich’]]

As prefixes and verb particles form a natural boundary within morphological analyses, the combinatorial path has to be pruned. For instance, if *Benutzerunterstützung* ‘user support’ is analyzed as in (10) - other analyses are possible - *unter* marks a boundary. After building all combinations for each of the subsets $\{$*’Be’ ’nutz’ ’er’*$\}$ and $\{$*’unter’ ’stütz’ ’ung’*$\}$, the Cartesian product of the resulting combinations has to be produced. The final sets of morphs and morph combinations are input for the weighting procedures.

(10)

Be	nutz	er	unter	stütz	ung
VPREF	V	NNSUFF	VPART	V	NNSUFF
be.pref	use	er.suff	below	support	ung.suff

4.2 Geometric Mean Score

Cap (2014, 67) uses the geometric mean as a quality measure for the analysis of German compounds. She uses the logarithmic transformation which is based the model of Koehn and Knight (2003). We use the non-transformed geometric mean as in (11) as the log-transformation preserves the ordering of the non-transformed value.

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad for \quad x_i ... x_n, \tag{11}$$

For the morph analysis of *Anbaumenge* to *An*|*bau*|*Menge* the respective morph frequencies are x_1 = 845 for *an*, x_2 = 168 for *bau* and x_3 = 8 for *Menge*, resulting in a *gm* score of 104.33. However, it is possible that the analyzed part *An*|*bau* is actually wrong and *Anbau* is the smallest unit that could be found. The same could hold for an analysis to *An*|*Baumenge*. However, the frequencies for these alternatives are lower than those of the first analysis (see Table 2).

4.3 Data Sparsity

The last example showed a case where a low frequency was consistent with linguistic reality. If the frequency of an element, whether a simple morph or an amalgamated form, is very small or 0, this can have two reasons: a. the form does not exist, or b. the form exists but is not present in the lexicon or in the underlying corpus. For example, the analysis of *10-Jahres-Prognosen* into its three lexical morphs would be impossible as numbers are not included in the lexicon. In both cases, the geometric mean would be undefined. However, especially for the second case, it is sensible to assign a small value to the element. Here, we chose 0.1. For a set of analyses which consists exclusively of unknown parts, this has the effect that the analysis with the smallest number of elements is chosen. This heuristic filters out longish pseudo-analyses which consist of highly frequent short words.

4.4 Heuristics for Parts of Speech

As surface and lexical forms of the two-level morphology might differ, we look up each morph or constituent candidate in both representations. For morphs which are the first part of the analysis, the lower case version has to be looked up, the opposite is necessary for nouns whose lexical form is represented with upper case, while their surface form might have lower case, if the noun is a non-initial component. SMOR produces the infinitive as the output for verbal morphs on the lexical level. However, for noun derivations with suffixes, the surface form of the verb stem is more relevant. After hyphens, the surface forms can start with a capital letter. Still other restrictions hold for abbreviations. A simple look-up heuristic deals with these different conditions.

5 Outcome and Evaluation

The following evaluation comprises qualitative and quantitative parts for each of the lexicons used. In the qualitative part, we consider cases of non-trivial analyses. The quantitative part presents results in terms of recall against the gold standard.

The qualitative test set covers three problems of disambiguation: a. ambiguity of morph boundaries, b. unknown parts of the analysis and c. ambiguous word structure.

For a. we choose the word forms

- *Anbaumenge* with the analyses *An|bau|menge* ‘(at|build|amount)’ and **An|Baum|Enge* ‘(at|tree|narrowness)’
- *Benzinverbrauch* with the analyses *Benzin|ver|brauch* ‘(petrol|(PREF)|use)’ and **Benzin|Verb|Rauch* ‘(petrol|verb|smoke)’
- *Aufbewahrungsorten* with the analyses of the immediate constituents **Aufbewahrung|sorte* ‘(storage|class)’ and *Aufbewahrung|s|orte* ‘(storage|(FL)|places)’; *Aufbewahrung* as a derived form can be analyzed as a complex multi-prefixed and suffixed form.

For b. we choose

- *10-Jahres-Prognosen* with the analysis ‘(10|-|Jahr|es-|Prognose)’ where *10* is unknown.

As an example for c., ambiguous structures, we choose

- *Arzneimittelverkaufs* with the noun constituents (*Arznei|mittel|verkauf)* ‘(medicine| means|sale)’. However, the next level of the morphological tree could either be *((Arznei|mittel)|verkauf)* ‘(medicine means|sale)’ or *(Arznei|(mittel|verkauf))* ‘*(medicine|means sale)’

For the quantitative evaluation, 50 percent (initial letters A to L) of the output of the system across the testset was evaluated by two humans. The data comprises 1,290 analyses of 572 word forms. These analyses are the ones representing the compositions with the highest score for a given item. No lower-ranking analyses are taken into account. For these analyses with largest scores, we annotated three cases:

* for wrong segmentations (false positive)

? for segmentations which were correct but on the “wrong level” (meaning higher-constituent analyses for morph analyses, or morph analyses instead of constituent analyses) (weak positive)

- for a correct segmentation (true positive, Recall)

We only considered the segmentation of the strings and ignored dubious tag assignments. However, if two analyses for the same word form got the same highest score and one of them was wrong, we marked this with *.

5.1 Morph Frequencies

5.1.1 Qualitative Analysis

Table 2 presents the output of the analyses. The morph analysis of *Anbaumenge* yields five different (flat) analyses from SMOR which can be combined into 16 plausible complex constructions. For each of the five SMOR results, the combinatorial analysis with the highest *gm* score is chosen.

word	gm	# of analyses	lexical analysis	surface analysis	tag structure
Anbaumenge	104.33	2	an|bauen|Menge	An|bau|menge	(VPART)(V|NNSUFF)(NN)
	104.33	4	an|bauen|Menge	An|bau|menge	(VPART)(V)(NN)
	12.66	4	an|baumen|eng	An|baum|eng	(VPART)(V)(ADJ|NNSUFF)
	9.19	4	an|baumenEnge	An|baumenge	(VPART)(V|NN)
	0.89	2	Anbau|Menge	Anbau|menge	(NN)(NN)
Benzinverbrauch	12.00	4	Benzin|Verb|Rauch	Benzin|verb|rauch	(NN)(NN)(NN)
	0.63	2	Benzin|Verbrauch	Benzin|verbrauch	(NN)(NN)
Aufbewahrungsorten	9.35	2	auf|bewahrenung|Sorte	Auf|bewahrung|sorte	(VPART)(V|NNSUFF)(NN)
	15.53	4	auf|bewahrenung|s|Ort	Auf|bewahrung|s|ort	(VPART)(V|NNSUFF|FL)(NN)
	8.25	4	auf|bewahrenungsorten	Auf|bewahrungsorten	(VPART)(V|NNSUFF|FL|V|NNSUFF)
10-Jahres-Prognosen	2.56	4	10|-|Jahr|es-|Prognose	10|-|Jahr|es-|Prognose	(PREF|HYPHEN)(NN|FL|HYPHEN)(NN)
Arzneimittelverkaufs	80.52	4	Arznei|Mittel|ver|kaufen	Arznei|mittel|ver|kauf	(NN)(NN)(VPREF)(V|NNSUFF)
	3.35	4	Arznei|Mittel|verkaufen	Arznei|mittel|verkauf	(NN)(NN)(V|NNSUFF)
	3.35	4	Arznei|Mittel|Verkauf	Arznei|mittel|verkauf	(NN)(NN)(NN)
	56.01	4	Arznei|mittel|ver|kaufen	Arznei|mittel|ver|kauf	(NN)(ADJ)(VPREF)(V|NNSUFF)
	2.07	4	Arznei|mittel|verkaufen	Arznei|mittel|verkauf	(NN)(ADJ)(V|NNSUFF)
	62.07	4	Arznei|mittel|Verkauf	Arznei|mittel|verkauf	(NN)(ADJ)(NN)
	22.36	2	Arzneimittel|ver|kaufen	Arzneimittel|ver|kauf	(NN)(VPREF)(V|NNSUFF)
	0.10	2	Arzneimittel|verkaufen	Arzneimittel|verkauf	(NN)(V|NNSUFF)
	0.10	2	Arzneimittel|Verkauf	Arzneimittel|verkauf	(NN)(NN)

Table 2: Output of morph analyses with *gm* score, number of compositions, lexical analysis, surface analysis and tag structure

It can easily be seen that the wrong analysis with *An|baum|enge* in the third line has a far lower score than the correct analysis. Another analysis based on the verb *baumen* ‘to sit on a tree’ and *Enge* ‘narrowness’ also gets a low score. The immediate constituents have a very low score, as *Anbau* is not part of the set of known morphs and only gets the back-off value of 0.1. The output for *Benzinverbrauch* faces the problem that SMOR does not segment the derived form *Verbrauch*. However, this word form is not part of the morph lexicon so that the analysis wrongly gives the best score to **Benzin|Verb|Rauch* ‘(petrol|verb|smoke)’ For *Aufbewahrungsorten*, the correct SMOR analysis gets the highest score. However, the score for the incorrect analysis **Auf|bewahrungsorte* ‘(VPART| keeping class)’ is surprisingly high, which is due to the high frequency of the verb particle *auf* which is multiplied by the sparse data value 0.1. The word form with the unknown number, *10-Jahres-Prognosen*, is correctly analyzed out of four different compositions. Due to the sparse data value, the *gm* score for each of these compositions can be calculated and compared. Finally, the example for ambiguous structures *Arzneimittelverkaufs* yields 9 SMOR analyses with 30 plausible combinatorial analyses. As can be seen from the last block of Table 2, the correct morph analysis gets the highest score. Note that the segmentation in line four with the second-largest score is also correct. However, it is based on an incorrect POS-assignment: *mittel* is analyzed as an adjective (‘middle’) instead of a noun.

5.1.2 Quantitative Analysis

For 572 word forms, we found 38 wrong segmentations and 70 cases which were correct annotations, though not on the expected morphological level. This leads to a recall of 81.11 percent. The number of different combinatorial analyses available was not taken into account.

About a third of the incorrectly analyzed word forms are of the type *An|passungsmechanismus*, where a high-frequency prefix determines the high score of a mostly unanalyzed word form.

Regarding the weak recall, some morph analyses are simply not feasible as the SMOR output does not always yield the smallest lexical units.

5.2 Frequencies of Constituents

5.2.1 Qualitative Analysis

The constituent analysis of *Anbaumenge* yields the same best analysis as the morph analysis. The numbers are slightly different but the score for the segmentation *Anbau|menge* is outweighed by that for *An|bau|Menge* due to the high frequencies of *an* and *bau*. The first part of Table 3 presents the output of the analyses.

The output for *Benzinverbrauch* is shown in the second part of Table 3. As with the morph-

<table>
<tr><th>word</th><th>gm</th><th># of analyses</th><th>lexical analysis</th><th>surface analysis</th><th>tag structure</th></tr>
<tr><td>Anbaumenge</td><td>59.80</td><td>2</td><td>an|bauen|Menge</td><td>An|bau|menge</td><td>(VPART)(V|NNSUFF)(NN)</td></tr>
<tr><td></td><td>53.70</td><td>4</td><td>an|bauen|Menge</td><td>An|bau|menge</td><td>(VPART)(V)(NN)</td></tr>
<tr><td></td><td>8.54</td><td>4</td><td>an|baumen|eng</td><td>An|baum|eng</td><td>(VPART)(V)(ADJ|NNSUFF)</td></tr>
<tr><td></td><td>6.05</td><td>4</td><td>an|baumenEnge</td><td>An|baumenge</td><td>(VPART)(V|NN)</td></tr>
<tr><td></td><td>4.90</td><td>2</td><td>Anbau|Menge</td><td>Anbau|menge</td><td>(NN)(NN)</td></tr>
<tr><td>Benzinverbrauch</td><td>6.51</td><td>4</td><td>Benzin|Verb|Rauch</td><td>Benzin|verb|rauch</td><td>(NN)(NN)(NN)</td></tr>
<tr><td></td><td>4.00</td><td>2</td><td>Benzin|Verbrauch</td><td>Benzin|verbrauch</td><td>(NN)(NN)</td></tr>
<tr><td>Aufbewahrungsorten</td><td>5.30</td><td>2</td><td>auf|bewahrenung|Sorte</td><td>Auf|bewahrung|sorte</td><td>(VPART)(V|NNSUFF)(NN)*</td></tr>
<tr><td></td><td>12.10</td><td>4</td><td>auf|bewahrenung|s|Ort</td><td>Auf|bewahrung|s|ort</td><td>(VPART)(V|NNSUFF|FL)(NN)</td></tr>
<tr><td></td><td>6.11</td><td>4</td><td>auf|bewahrenungsorten</td><td>Auf|bewahrungsorten</td><td>(VPART)(V|NNSUFF|FL|V|NNSUFF)</td></tr>
<tr><td>10-Jahres-Prognosen</td><td>2.29</td><td>4</td><td>10|-|Jahr|es-|Prognose</td><td>10|-|Jahr|es-|Prognose</td><td>(PREF|HYPHEN)(NN|FL|HYPHEN)(NN)</td></tr>
<tr><td>Arzneimittelverkaufs</td><td>43.4</td><td>4</td><td>Arznei|Mittel|ver|kaufen</td><td>Arznei|mittel|ver|kauf</td><td>(NN)(NN)(VPREF)(V|NNSUFF)</td></tr>
<tr><td></td><td>12.80</td><td>4</td><td>Arznei|Mittel|verkaufen</td><td>Arznei|mittel|verkauf</td><td>(NN)(NN) (V|NNSUFF)</td></tr>
<tr><td></td><td>12.80</td><td>4</td><td>Arznei|Mittel|Verkauf</td><td>Arznei|mittel|verkauf</td><td>(NN)(NN) (NN)</td></tr>
<tr><td></td><td>29.50</td><td>4</td><td>Arznei|mittel|ver|kaufen</td><td>Arznei|mittel|ver|kauf</td><td>(NN)(ADJ)(VPREF) (V|NNSUFF)</td></tr>
<tr><td></td><td>7.65</td><td>4</td><td>Arznei|mittel|verkaufen</td><td>Arznei|mittel|verkauf</td><td>(NN)(ADJ) (V|NNSUFF)</td></tr>
<tr><td></td><td>7.65</td><td>4</td><td>Arznei|mittel|Verkauf</td><td>Arznei|mittel|verkauf</td><td>(NN)(ADJ) (NN)</td></tr>
<tr><td></td><td>10.6</td><td>2</td><td>Arzneimittel|ver|kaufen</td><td>Arzneimittel|ver|kauf</td><td>(NN)(VPREF) (V|NNSUFF)</td></tr>
<tr><td></td><td>0.84</td><td>2</td><td>Arzneimittel|verkaufen</td><td>Arzneimittel|verkauf</td><td>(NN)(V|NNSUFF)</td></tr>
<tr><td></td><td>0.84</td><td>2</td><td>Arzneimittel|Verkauf</td><td>Arzneimittel|verkauf</td><td>(NN)(NN)</td></tr>
</table>

Table 3: Output of constituent analyses with gm-score, number of compositions, lexical analysis, surface analysis and tag structure

based analysis, the constituent analyses wrongly gives the best score to **Benzin|Verb|Rauch* (petrol|verb|smoke). The respective frequencies of the constituents within the CELEX lexicon are (4, 3, and 23) vs. (4 and 4), so the segmentation into three parts is preferred. The third part of Table 3 presents the results for *Aufbewahrungsorten*. While the highest rank remains the same, there is an increase in the scores as the immediate constituent counts increment the number of longer words at the cost of the shorter ones. The analysis of *10-Jahres-Prognosen* is the same as for the morphs, as all constituents are monomorphemes. Due to the different counts, the *gm* score differs slightly. The analysis of ambiguous structures for *Arzneimittelverkaufs* can be seen in the last part of Table 3. Though closer to each other, the scores are in the same order as for the morph analyses. This unwanted result is caused by the low frequency for the constituent *Arzneimittel* in CELEX.

5.2.2 Quantitative Analysis

We found 22 wrong segmentations and 86 weak positive ones. The word forms concerned were the same as for the morphs, which results in the same overall recall. Sometimes the sequence of the *gm* scores was a bit closer to the correct order, however this frequency count shows that constituent counts from a lexicon of an acceptable size are not good enough for analyzing these cases. It is of some linguistic irony that the weak positive annotated analyses are mostly good analyses for the morph level or another low-level description, e.g. the *Fuß|ball|national|team* ‘national football team’ was correctly analyzed.

Among the wrongly analyzed forms we encounter the above-described effect of too dominant prefixes. Segmentations such as **Benzin|Verb|Rauch* are rather an exception.

5.3 Corpus Frequencies

5.3.1 Qualitative Analysis

As the frequency counts for the corpora are higher than those for the lexicons, the scores also become larger and tend to differ more significantly. Table 4 shows the results for *Anbaumenge*. Obviously, the ranks are determined by the high frequencies of prefixes and verb particles.

Benzinverbrauch is analyzed correctly, though the tag structure reveals that the analysis was produced by merging *Verb* with *rauch* to *Verbrauch*. *Aufbewahrungsorte* and *10-Jahres-Prognosen* yield good results too. *Arzneimittelverkaufs* is perfectly analyzed for the morph level. In general, the corpus-based analyses of constituents and morphs produce more interpretable results.

gm	analyses	lexical analysis	surface analysis	tag structure
6075.87	2	an|bauen|Menge	An|bau|menge	(VPART)(V|NNSUFF)(NN)
6075.87	4	an|bauen|Menge	An|bau|menge	(VPART)(V)(NN)
209.14	4	an|baumen|eng	An|baum|enge	(VPART)(V)(ADJ|NNSUFF)
85.27	4	an|baumenEnge	An|baumenge	(VPART)(V|NN)
114.60	2	Anbau|Menge	Anbau|menge	(NN)(NN)

Table 4: Output of word-form analysis for *Anbaumenge* with gm-score, number of compositions, lexical analysis, surface analysis and tag structure

5.3.2 Quantitative Analysis

Scores for the word-form frequencies differed from the lexical frequencies and the results were improved. The errors that remain include syntagmatic compounds such as *50-jährig* which are erroneously segmented as endocentric compounds, e.g. *50*|-|*jährig* '50|-|year+suffix'. Even ambiguous forms on the morph level (*Benzinverbrauch*) are segmented correctly on the string level – though their morphological analyses remain erroneous.

Table 5 gives an overview of the evaluation with the overall recall and the recall for the weakly consistent analyses.

dataset	*	?	overall	weak
			recall	recall
morphs	38	70	81.11	93.34
constituents	22	86	81.11	96.15
word forms	15	88	88.02	97.38

Table 5: Overall recall and weak recall for three frequency sources

6 Discussion

The approach we presented shows how different frequency counts can lead to different specific constituent segmentation analyses. The corpus frequencies in particular lead to better segmentation on the morph level.

Ideally, we would derive counts from corpus data that match the register and domain of the lexical units that are to be analyzed. Here, frequencies derived from a corpus of 6.0 million words (Gulikers et al., 1995, 102) were too small to yield reliable counts for non-monomorphic constituents. Larger sets of well-annotated corpus data should be used. Moreover, the analysis process can be augmented by other linguistic characteristics: parts of speech, position of constituents in words, and the text specificity of words.

As our approach builds on the output of a morphological segmentizer, it is dependent on the prior segmentation, for better or for worse. Starting with analyses of different morphological tools might help to avert, or compensate for, gaps in the lexicon or systematic weaknesses in the tools. In general, morphological data should be analyzed from many sides.

The use of the geometric mean should be analyzed from a quantitative point of view. In particular, the distribution of values should be investigated to make statements about their relevance. When weighting alternatives of n vs. $n + 1$ constituents, in most cases the geometric mean for the variant with more constituents is larger than that for the variant with fewer constituents. This is owed to the facts that a) shorter morphs and lexical units are more frequent than longer ones and b) corpus frequencies for compounds or derivates are still relatively small compared to the frequencies of their constituents. Restricting the cotext to smaller units than the corpus, such as the document or paragraph, could help, although in that case the data might become too sparse.

7 Conclusion

This investigation has shown that ambiguous flat structure results from morphological analyses can be disambiguated by using additional statistical methods, especially on the morphological level.

What could not be analyzed on a lower level should be re-analyzed by using the same combinatorial approach on a higher-level analysis. The current state of work already shows some more complex morphological structures. However, as the geometric mean is not a good indicator for the concatenation of morphs, the weighting measure(s) should be derived carefully. In future work, we will explore probabilistic models. In combination with such models, the sets of integer compositions for the analyses of one word form can be processed in transition networks. Also, we will focus on building up more levels of the word-structure trees and exploiting the statistical dependencies between morphs and their parts of speech.

Acknowledgments

The authors were partially supported by the German Research Foundation (DFG) under grant RU 1873/2-1. We would like to thank especially Helmut Schmid for his valuable advice and updates of the SMOR lexicons and Jasper Brandes for his help in the evaluation task.

References

Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. The CELEX lexical database (CD-ROM).

Fabienne Cap. 2014. *Morphological processing of compounds for statistical machine translation*. Ph.D. thesis, Universität Stuttgart.

Alexander Geyken and Thomas Hanneforth. 2006. TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002, pages 55–66. Springer.

Léon Gulikers, Gilbert Rattink, and Richard Piepenbrock. 1995. German Linguistic Guide. In Harald Baayen, Richard Piepenbrock, and Léon Gulikers, editors, *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, Philadelphia, PA.

Mariikka Haapalainen and Ari Majorin. 1995. GERTWOL und morphologische Disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference on Computational Linguistics, Helsinki, Finland*.

Gerhard Hanrieder. 1991. Robustes Wortparsing. Lexikonbasierte morphologische Analyse (komplexer) deutscher Wortformen. Master's thesis, Universität Trier.

Gerhard Hanrieder. 1996. MORPH - Ein modulares und robustes Morphologieprogramm für das Deutsche in Common Lisp. In Roland Hauser, editor, *Linguistische Verifikation Dokumentation zur Ersten Morpholymics 1994*, pages 53–66. Niemeyer, Tübingen.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 187–193. Association for Computational Linguistics.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *Proceedings of the 10th international conference on Computational linguistics*, pages 178–181. Association for Computational Linguistics.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 1063–1067.

Kay-Michael Würzner and Thomas Hanneforth. 2013. Parsing morphologically complex words. In Mark-Jan Nederhof, editor, *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, FSMNLP 2013, St. Andrews, Scotland, UK, July 15-17, 2013*, pages 39–43. The Association for Computer Linguistics.