Distinguishing affixoid formations from compounds

Josef Ruppenhofer
LeibnizScienceCampus LiMo
Institute for German Language
Mannheim
ruppenhofer@ids-mannheim.de

Michael Wiegand
Spoken Language Systems
Saarland University
michael.wiegand@lsv.uni-saarland.de

Rebecca Wilm
Katja Markert
LeibnizScienceCampus LiMo
Computational Linguistics
Heidelberg University
{wilm,markert}@cl.uni-heidelberg.de

Abstract

We study German affixoids, a type of morpheme in between affixes and free stems. Several properties have been associated with them – increased productivity; a bleached semantics, which is often evaluative and/or intensifying and thus of relevance to sentiment analysis; and the existence of a free morpheme counterpart – but not been validated empirically. In experiments on a new data set that we make available, we put these key assumptions from the morphological literature to the test and show that despite the fact that affixoids generate many low-frequency formations, we can classify these as affixoid or non-affixoid instances with a best F1-score of 74%.

1 Introduction

In this work, we study a subset of complex German words that includes many hapaxes, namely items that have as one of their morphological components a so-called affixoid or semi-affix. Examples of affixoid formations include Gesetzhengst ‘jobsworth, stickler for the letter of the law’ [lit. ‘law/legal stallion’] or Mentalitatsmonster ‘person with laser focus’ [lit. ‘mentality monster’]. The class of affixoid morphemes sits in between affixes and stems. While various criteria have been proposed to identify affixoids (Schmidt, 1987), at least the following three are widely taken for granted (ten Hacken, 2000; Elsen, 2009): (i) increased productivity; (ii) semantic bleaching/decreased semantic specificity; and (iii) an etymological and formal link to an existing free stem. The first two criteria are applied by comparing the affixoid to its corresponding free stem. For instance, in Weingott ‘deity of wine’, Gott is a free morpheme occurring with its regular meaning; in Gitarregott ‘guitar god’, Gott is an affixoid: the whole word does not refer to a deity but to a talented human guitar player. Such evaluative or intensifying meanings are a hallmark of affixoids (Meibauer, 2013). The last of the three criteria distinguishes affixoids from affixes, which by definition occur only bound to other morphemes. For example, German -heit, whose English cognate is -hood as in falsehood, is an affix: there is no longer a related free form. By contrast, German -gott is an affixoid since there is a related free form Gott, cognate with English god.

Existing research on affixoids is centered on Germanic languages like German, Dutch and Swedish (Ascoop and Leuschner, 2006; Booij, 2005; Booij and H¨uning, 2014; Norde and Van Goethem, 2014). However, we believe affixoids are not an exclusive feature of these languages. They are likely to arise in other languages in which compounding is productive. For English, for instance, there is little to no systematic research but arguably quality (as in quality press/furniture/diamonds but not in quality management) and nut (as in health/math/trivia nut but not in pecan nut) can be considered English prefixoids and suffixoids, respectively. Even for languages on which there is more research, that work is typically focused on the theoretical relevance of assuming a category of affixoids that is distinct from affixes on the one hand and compounds on the other. Very little quantitative work using corpus data has been done to, for instance, study the level of productivity for different affixoid candidates or to substantiate the intuition that most affixoid uses carry evaluative meanings. Our work thus wants to fill an empirical gap in the theoretical discussion.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/
In addition, we are interested in studying affixoids for the purposes of sentiment analysis. On the one hand, theoretical linguistic work that notes the expressive function of affixoids such as (Meibauer, 2013) suggests this. On the other hand, it is known from prior research on sentiment analysis that hapax words in general are often subjective (Wiebe et al., 2004). As we show in §3, affixoids tend to generate many (near-)hapax forms, which meshes with observations on their productivity in the morphological literature and the general expectation about the Zipfian distribution of word frequencies. Since hapaxes, by definition, cannot be readily analyzed based on their distribution in corpora, it would be very useful if we could make use of their intrinsic properties to classify such forms as affixoid uses (and therefore likely subjective) or not.

The main task that we set ourselves in this paper is morpheme sense disambiguation: we want to classify complex forms containing possible affixoids as to whether the morphemes in question really occur as affixoids with special meanings or as regular morphemes with their full, common semantics. We frame this task as a binary classification problem. The major contributions of our paper are:

- a gold standard annotation of complex forms containing potential affixoids, which is also going to be made publicly available;
- empirical validation of claims regarding the association between intensification / evaluation and affixoid meanings;
- a detailed examination of various novel features that have been devised to detect the presence of an affixoid in a complex form.

2 Related work

Recently, Ruppenhofer et al. (2017) studied the problem of how well the polarity of complex words, including both compounds and derived forms, can be predicted based on its components and the word’s morphological structure. They found that, while on core vocabulary – defined as items listed in the PolArt dictionary (Klenner et al., 2009) and highly likely to be listed in GermaNet (Hamp and Feldweg, 1997), the German WordNet resource – classification accuracy was as high as 85%, performance was severely degraded on more colloquial, domain-specific and linguistically creative forms taken from Wiktionary and the Wortwarte neologism project. Note that this research did not distinguish true compounds from complex forms involving affixoids.

Similar research was carried out earlier by Moilanen and Pulman (2008) who evaluated how well it was possible to classify unknown English words into one of three polarity classes based on morphological analysis. Compared to these prior efforts, our task is narrower as we do not deal with derivation and do not predict polarity for the complex entries. Our data is also more focused since we have only complex forms containing exactly two nouns and the second components of our data represent only 7 different lemmas. And importantly, our data are low frequency words in distinction to Ruppenhofer et al. (2017), whose main data set looked at higher-frequency core vocabulary.

In another line of research, several studies have tackled the problem of classifying senses as either objective or subjective. For English Su and Markert (2009) and Gyamfi et al. (2009) tackled this task on data from WordNet. Subjective entries are ones that possess polarity, that is, they could in further analysis be classified as either positive or negative. This task is similar to ours in the sense that a simple binary categorization of items is sought. However, there are key differences. First, our classification targets lemmas not word senses. That is, we assume that the complex lemmas are monosemous or have a clearly dominant sense in our data. Second, the distinction objective versus subjective does not fully line up with the distinction non-affixoid use versus affixoid use. A complex form may contain the affixoid candidate in its regular objective sense but the other component may make the whole word subjective. An example of this kind is Lieblingshai ‘favorite shark’. Finally, since our complex forms are unlikely

\footnote{http://wortwarte.de/}

\footnote{This assumption can be made plausible by the observations that longer words have more specific meanings than shorter ones, and that they tend to have fewer meanings than their head words, and fewer meanings than their components do on average (Altmann, 2002).}
to be listed in lexical resources, we will usually lack information such as glosses, supersenses or example sentences for them.

In another strand of research involving German morphology and sentiment analysis, Wiegand et al. (2016) developed an approach to classify the first element of German compounds as expressing either the source or target of evaluation, or neither, relative to the second element, if in the first step of analysis the second element was determined to be subjective. As do we, those authors focused on noun-noun compounds and they did not address polarity classification. However, their approach targets higher frequency words as it relies on the availability of sufficient corpus data to enable the use of distributional similarity. For our dataset, we cannot directly model the distributional properties of our complex items.

3 Data

Data creation. German has both nominal affixoids – items that are related to nouns (e.g. Affen|tempo, ‘high speed’ [lit. ‘ape speed’]) – and adjectival affixoids – items that are related to adjectives (e.g. unheils|schwanger ‘ominous’ [lit. ‘pregnant with doom’])\(^3\). Here, we concentrate on nominal affixoids. Two subtypes of nominal affixoids exist, depending on the position of the nominal affixoid in the complex word. Affe, as in Affentempo, is a prefixoid as it occurs as the first item in the compound-like complex word, whereas Gött is a suffixoid as it occurs as the second item of complex formations such as Gitarrengott ‘guitar god’. To make the best use of our resources for annotation, we focus on suffixoids here as we believe that addressing suffixoids and prefixoids at the same time may not be helpful. Consider that for many suffixoid candidates, a suffixoid use is recognizable by the fact that the referent of the whole word is not a hyponym of the suffixoid in its basic meaning. For instance, a Kredithai ‘loan shark’ is not a shark. With prefixoids, this is usually different, as they tend to add evaluation or intensification but do not change the class of referent. A Scheißauto ‘shit car’ is still a car.

No pre-compiled resource is available that lists all suffixoids and the complex words in which they figure. We therefore first compiled a set of 60 possible suffixoids from the literature on morphology (Motsch, 2004; Erben, 2006; Elsen, 2009; Elsen, 2011; Fleischer et al., 2012) and then looked for complex forms containing them in the 1.7-billion deWaC web corpus (Baroni et al., 2009).\(^4\) We queried the corpus for complex forms whose lemma ended in one of the suffixoids and where the suffixoid was preceded by at least four letters. The latter condition was imposed to rule out complex forms where the first part is a simple prefix. Further, we allowed only for complex forms spelled as contiguous strings as standard German orthography does not allow open compounds. E.g. the German word for hammerhead shark cannot be spelled as Hammer Hai using two white space-separated tokens.

The extracted forms were then semi-manually filtered for remaining errors. We used the SMOR morphological segmentation tool (Schmid et al., 2004; Faaß et al., 2010) to detect cases that did not represent simple noun-noun combinations. We excluded instances that were not compliant with standard German orthography, for instance, due to spelling errors or incorrect tokenization. We eliminated duplicates that differed only in whether the two parts of the complex form were written contiguously or separated by a hyphen. Thus, we for instance kept Hammerhai but not Hammer-Hai ‘hammerhead shark’.

To keep the overall effort manageable, we selected seven items from our pool of suffixoids for annotation and subsequent use in our experiments. We chose these for two reasons. First, we did not want items that were extremely biased towards either affixoid or non-affixoid uses. For instance, more than 99% of the formations for Dreck ‘dirt’ and Junkie ‘junkie, addict’ are affixoid formations. Second, we wanted items for which a sizable number of complex forms exist in our corpus. This criterion ruled out, for instance, Base ‘female cousin’ because most forms found in the corpus were false positives involving compounds using the homographic English noun base as a head, rather than intended German forms such as Klatschbase ‘telltale/chatterbox’. Table 1 gives some basic descriptive information about the

---

\(^3\)Our literature survey did not throw up any German verbal affixoids. This is not unexpected since German also has very limited compounding involving verbs. They can only occur as modifiers in compounds with nominal or adjectival heads as in Bratpfanne ‘frying pan’ or waschecht ‘colorfast’ [lit. ‘wash-true’] (Olsen, 2000).

\(^4\)We also found 40 possible nominal prefixoids in the literature we reviewed. Note that these lists are not necessarily complete; the items discussed in the linguistic literature are discovered through introspection rather than in a data-driven way.
Table 1: Properties of the suffixoid candidates; senses = major senses as defined by duden.de; frequency in deWaC; frequency and frequency rank according to dlexdb.

<table>
<thead>
<tr>
<th>basic sense</th>
<th># senses</th>
<th>intensifying</th>
<th>polar</th>
<th>freq</th>
<th>freq. rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolzen</td>
<td>4</td>
<td>Y</td>
<td>Y</td>
<td>115</td>
<td>3462.5</td>
</tr>
<tr>
<td>Bruder</td>
<td>4</td>
<td>N</td>
<td>Y</td>
<td>12901</td>
<td>734.0</td>
</tr>
<tr>
<td>Gott</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>36488</td>
<td>245.0</td>
</tr>
<tr>
<td>Hai</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>247</td>
<td>19654.5</td>
</tr>
<tr>
<td>Hengst</td>
<td>2</td>
<td>N</td>
<td>Y</td>
<td>342</td>
<td>15551.5</td>
</tr>
<tr>
<td>Kaiser</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>12040</td>
<td>798.0</td>
</tr>
<tr>
<td>König</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>21439</td>
<td>415.0</td>
</tr>
</tbody>
</table>

The complex forms of these 7 items will represent the raw data for our gold standard (see also supplementary material). In total, there are 1788 of such complex forms.

Table 2 shows for each suffixoid candidate the number of complex forms in the gold standard annotations in which the candidate occurred as true suffixoid (Y), the number of complex forms in which it occurred as a regular compound head, cases where the annotators were unsure whether the candidates were mainly used as an affixoid or not, and the total of all forms found.

<table>
<thead>
<tr>
<th>basic sense</th>
<th>Y</th>
<th>N</th>
<th>both</th>
<th>unsure</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bolzen</td>
<td>30</td>
<td>70</td>
<td>4</td>
<td>1</td>
<td>105</td>
</tr>
<tr>
<td>-bruder</td>
<td>21</td>
<td>158</td>
<td>3</td>
<td>1</td>
<td>183</td>
</tr>
<tr>
<td>-gott</td>
<td>78</td>
<td>354</td>
<td>41</td>
<td>0</td>
<td>473</td>
</tr>
<tr>
<td>-hai</td>
<td>16</td>
<td>49</td>
<td>0</td>
<td>3</td>
<td>68</td>
</tr>
<tr>
<td>-hengst</td>
<td>26</td>
<td>81</td>
<td>3</td>
<td>1</td>
<td>111</td>
</tr>
<tr>
<td>-kaiser</td>
<td>31</td>
<td>92</td>
<td>11</td>
<td>0</td>
<td>135</td>
</tr>
<tr>
<td>-könig</td>
<td>290</td>
<td>370</td>
<td>53</td>
<td>0</td>
<td>713</td>
</tr>
<tr>
<td>total</td>
<td>492</td>
<td>1174</td>
<td>115</td>
<td>7</td>
<td>1788</td>
</tr>
</tbody>
</table>

Table 2: Affixoid and non-affixoid uses (gold)

<table>
<thead>
<tr>
<th>basic sense</th>
<th>dlexdb</th>
<th>GermaNet</th>
<th>Wiktionary</th>
<th>SentiMerge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>aff</td>
<td>all</td>
<td>aff</td>
</tr>
<tr>
<td>-bolzen</td>
<td>17 (16.5)</td>
<td>1</td>
<td>1 (1.0)</td>
<td>0</td>
</tr>
<tr>
<td>-bruder</td>
<td>41 (22.4)</td>
<td>4</td>
<td>3 (1.6)</td>
<td>0</td>
</tr>
<tr>
<td>-gott</td>
<td>133 (28.1)</td>
<td>10</td>
<td>21 (4.4)</td>
<td>2</td>
</tr>
<tr>
<td>-hai</td>
<td>13 (19.1)</td>
<td>3</td>
<td>7 (10.3)</td>
<td>1</td>
</tr>
<tr>
<td>-hengst</td>
<td>23 (20.7)</td>
<td>7</td>
<td>2 (1.8)</td>
<td>1</td>
</tr>
<tr>
<td>-kaiser</td>
<td>25 (18.5)</td>
<td>3</td>
<td>1 (2.8)</td>
<td>0</td>
</tr>
<tr>
<td>-könig</td>
<td>177 (24.8)</td>
<td>54</td>
<td>19 (2.6)</td>
<td>5</td>
</tr>
<tr>
<td>total</td>
<td>429 (24.0)</td>
<td>82</td>
<td>54 (3.0)</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3: Coverage by lexical resources (absolute and %)

Table 3 shows that only 429 items (24.0%) of the 1788 forms in the gold standard are covered by the dlexdb lexical database (Heister et al., 2011), 54 (3.0%) are covered by GermaNet 11.0, and only 28 (1.6%) by Wiktionary. The coverage by SentiMerge (Emerson and Declerck, 2014), the largest German sentiment lexicon with close to 100k entries, is also low. By contrast, 93% of the 9300 items in Ruppenhofer et al. (2017)'s main data set were covered by dlexdb, and 88.4% by GermaNet. Finally, Table 3 shows that of the formations included in the various resources, the majority are non-affixoid cases. Inclusion in a resource thus is not a very useful feature as it is only predictive of the majority class.

Figure ?? illustrates that the token frequencies for the complex forms with König ‘king’ as its second component have a largely Zipfian distribution: there are many complex forms with very low token frequencies (in fact, mostly their frequency is 1), whereas there are very few complex forms with high token frequencies. While we cannot show it here for lack of space, the shape of the distribution is the same for all the suffixoid candidates. Note further that the numbers shown in Table 2 do not delimit the productivity of these suffixoids. Consider that, for instance, in a collection of 120 million tweets, we found 905 complex forms for König, of which only 226 overlapped with the 713 forms found in the deWaC corpus.

**Annotation of the affixoid vs. non-affixoid distinction.** Most of the data was annotated by one of the
authors. To assess how well human annotators agree on the distinction between affixoid uses and non-affixoid uses, two of the authors performed an annotation experiment on 200 randomly chosen instances of the total set of 1788 complex forms. The annotators could label each complex form as Y (only affixoid uses or predominantly affixoid uses), N (only non-affixoid uses or predominantly non-affixoid uses) and B (both types of use likely). The annotators achieved a Cohen’s kappa (Cohen, 1960) value of 0.67, which amounts to substantial agreement according to Landis and Koch (1977). As the confusion matrix in Figure 4 shows, the majority of errors are ‘milder’ cases where one annotator considers one type of use dominant whereas the other considers both types of uses plausible.

### 4 Features

#### 4.1 Automatic features

We first explore features that can be extracted automatically from corpora and resources, which is the most realistic setting.

**Suffixoid.** Since the suffixoid candidates differ in how frequently they do in fact occur as affixoids and also differ in their evaluative meanings, it makes sense to capture which suffixoid in particular occurs in a complex form. We define a binary indicator variable for each suffixoid.

**Frequency.** There exists a significant body of research in quantitative linguistics on the productivity of morphological processes. But works on the productivity of compounding such as Altmann (2002) or Krott et al. (1999) do not make a distinction between affixoid formations and true compounds and therefore there are no prior results on their similarities and differences in regards to productivity: the assumption of increased productivity for affixoids has, to our knowledge, not been empirically validated. For our set suffixoids, we used Baayen’s `languageR` R-package to compare, the lexical richness of compound formations to that of affixoid formations, per affixoid candidate and for pooled data. We found no consistent differences in terms of vocabulary growth curves, vocabulary growth rates or type-token ratios (Baayen, 2005). This result notwithstanding, we want to see if frequency information about the components and the complex forms could help us distinguish between the two kinds of complex forms. One relevant intuition that we had in this regard is that the complex forms of affixoidal uses may tend to have lower frequencies than complex forms representing regular compounds. Table 5 shows that for most items this seems to be borne out, although the difference is not statistically significant, with the exception of -hai. However, for -könig the reverse situation holds in a statistically significant way: the frequencies of affixoid formations are higher than those of the non-affixoid formations.

<table>
<thead>
<tr>
<th>affixoid</th>
<th>non-affixoid</th>
<th>p-value of t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>-hai</td>
<td>37.50</td>
<td>12.41</td>
</tr>
<tr>
<td>-kaiser</td>
<td>3.40</td>
<td>6.64</td>
</tr>
<tr>
<td>-bolzen</td>
<td>4.21</td>
<td>4.36</td>
</tr>
<tr>
<td>-hengst</td>
<td>4.03</td>
<td>8.40</td>
</tr>
<tr>
<td>-könig</td>
<td>4.92</td>
<td>16.23</td>
</tr>
<tr>
<td>-bruder</td>
<td>1.88</td>
<td>6.04</td>
</tr>
<tr>
<td>-gott</td>
<td>10.07</td>
<td>19.90</td>
</tr>
</tbody>
</table>

Table 5: Frequency differences between affixoid and non-affixoid formations

**Compositionality.** This feature measures the compositionality of the complex word. The intuition is as follows. In regular compounds, the meaning of the complex form is more or less compositionally derived from the meaning of the components. However, since affixoid uses of the morphemes in question go along with bleached meanings, the semantics of a complex word containing an affixoid use should
be harder to model compositionally. Following Schulte im Walde et al. (2013), we represent each com-
ponent as well as the complex word as vectors. As our measure of compositionality, we determine the
cosine similarity between the sum of the component vectors and the vector of the complex word.

We used fastText (Bojanowski et al., 2016) in its default setting to train word embeddings on the
SdeWaC-corpus (Faaß and Eckart, 2013), which contains about 880 million tokens and is a cleaned up
version of the deWaC corpus (Baroni et al., 2009). The choice of fastText was motivated by the fact that
fastText computes vectors for words by adding up the vectors for n-grams found in the words, which
allows us to produce vectors for words not seen in the training data. Since many of our complex forms
are (near-)hapaxes, this is a crucial benefit of fastText.

**Pointwise mutual information.** We use Pointwise mutual information (Church and Hanks, 1990) to
capture the level of association between the two components of the complex word. The expectation is
that the components of regular compounds exhibit higher PMI-scores than the components of a complex
word involving an affixoid. This is motivated by Tellenbach (1985)’s observation that for complex forms
containing an affixoid use, paraphrases containing the morpheme in question as a free word are unlikely.
By contrast, compositional compounds are often paraphrased using their components. As an example, the
compositional compound *Perserkönig* is readily paraphrased as *König der Perser* ‘king of the Persians’,
whereas *Gitarrenkönig* is less likely to be paraphrased as *König der Gitarre* ‘king of the guitar’. Even
aside from paraphrasing, it seems more likely that *Perse* and *König* occur together in a text as they share
the context of governance than it does that *Gitarre* and *König* co-occur.

**GermaNet supersenses.** As in the English WordNet (Miller, 1995), GermaNet’s synsets are associ-
ated with supersenses that define high-level categorizations such as Human, Animal, Artifact, or Group. A lexical unit in GermaNet can be associated with one or more of its supersenses. Accordingly,
we define three series of indicator variables that capture whether any sense of the complex form, the first
component or the second component exhibits a particular supersense. We refer to all these features as
supersenses_bag. This feature group is motivated by the observation that with several of our suffix-
oid candidates, the semantic types of the first component and the complex form tends to provide good
evidence on whether the complex word will be an affixoidal use or not. For instance, the first component
*Haflinger* in *Haflinger|hengst* refers to a breed of horses and the complex form is a regular compound.
By contrast, the first component *Büro* ‘office’ of *Büro|hengst* refers to either an Artifact or a Group
and the complex form represents an affixoidal use meaning something like ‘pencil pusher’. Moreover,
the complex form *Haflinger|hengst* also has the supersense Animal, whereas *Büro|hengst* has the sup-
sersense Human. Because for this particular set of suffixoid candidates differences between the second
component and the complex word’s supersenses may be particularly important, we experiment with an
alternative set of supersense features (supersenses_diffs): we use a series of indicator variables that code
whether the second component and the complex word differ in their value for a given supersense.\(^5\)

**Polarity.** Since affixoid uses are likely to have evaluative meanings, we explore whether this is re-
flected in the polarity of the two components and the complex form. We extract polarity information for
all three from SentiMerge (Emerson and Declerck, 2014). With 96,918 entries, it is to date the largest
available polarity lexicon for German. SentiMerge was created by harmonizing and combining three
smaller lexicons (PolArt (Klenner et al., 2009); GermanPolarityClues (Waltinger, 2010); and SentiWS
(Remus et al., 2010)) using a Bayesian probabilistic model.

**Psycholinguistic features.** If available, we extract psycholinguistic ratings along four dimensions
for the whole word and its components. This type of feature has been successfully used in various
tasks, such as identifying metaphors (Turney et al., 2011; Klebanov et al., 2014); studying persuasion
(Tan et al., 2016); sarcasm detection (Bamman and Smith, 2015); and, most similar to us, polarity
prediction for complex words (Ruppenhofer et al., 2017). The first dimension places words on a scale
from abstract to concrete (abstconc). Abstract words refer to things that we cannot perceive directly with
our senses (integer, politics, ...) whereas concrete words refer to things we can perceive (sound, scent,
...). The second dimension concerns imageability (img). A large subset of concrete words have a high

\(^5\)Note that for other suffixoids not covered here such as *Papst* (lit. ‘pope’, suffixoid ‘expert’) and *Nest* (lit. ‘nest’, affixoid
‘den/hideout’) there is no difference at all between the supersenses of the second component and the complex word.
imageability. These words refer to things that we can actually see (chestnut, police jeep, ...). The third rating dimension is valence (val), which measures the pleasantness of a word (gift vs. punishment). The final dimension, arousal, represents the intensity of emotion caused by a stimulus (alert vs calm). We obtain affective ratings from the resource of Köper and Schulte im Walde (2016). It provides information on 350k words and is far more comprehensive than the affective norm data of Kanske and Kotz (2010) or Lahl et al. (2009). It is also much larger than commonly used polarity lexicons for German such as PolArt (Klenner et al., 2009) or GermanPolarityClues (Waltinger, 2010).

Emotion. Since emotion information is commonly used in sentiment-related classification tasks (e.g. Tang et al. (2014), Sulis et al. (2016)), we wanted to see to what extent emotion information could benefit our task. For this purpose, we use the NRC Word-Emotion Association Lexicon (EmoLex) for English which was created by Mohammad and Turner (2013) using a crowdsourcing approach. EmoLex contains binary associations of words with the eight basic emotions (joy, sadness, anger, fear, disgust, surprise, trust, anticipation) of Plutchik (1962).

Although the German version of the lexicon was produced using machine translation, we use it here because we do not have a similarly large natively produced resource available for German. The German EmoLex covers 9630 lemmas. For each complex form, we extract the emotions associated with the overall word and do likewise for the first and the second components.

4.2 Missing value imputation

With features derived from resources (polarity, psycholinguistic ratings, emotion, supersenses and pmi), we face the problem that there are gaps in coverage. Various strategies are conceivable to fill in missing values. The first one we considered is to substitute for a word that is not covered that word which is most similar to it according to cosine similarity among the fastText vectors and which is covered by the resource. A second approach we considered is to use the mean value for the feature in question. A third approach uses the median. A fourth option is to use the modal value. Our experiments showed that for our SVM classifier, the choice of imputation strategy does not result in any statistically significant differences in results. We thus report results based on the first strategy, which is based on cosine similarity.

4.3 Manual feature annotations

In order to be able to explore to what extent the notions of intensification and evaluation correlate with the use of our target morphemes as either affixoids or regular morphemes, we added further annotations to our complex forms. We labeled them manually with respect to the following features.

- Polarity of complex word: does the item have positive, negative or neutral polarity?
- High intensity of complex word: does the complex word express high intensity? For Temperamentsbolzen ‘highly temperamental person’ this is true, but not for Metallbolzen ‘metal bolt’.
- Evaluativity of complex word: does the complex word mostly carry either a positive or negative evaluation? True for Bürohengst ‘pencil pusher’, but not for Haflingerhengst ‘Haflinger stallion’.
- Head Intensity: does the head of the complex word, that is the potential affixoid, contribute an intensifying meaning to the overall word? For Temperamentsbolzen ‘highly temperamental person’ is true but for Riesenbolzen ‘giant bolt’ it is false.
- Head Polarity: does the head of the complex word contribute an evaluating meaning to the overall word? For Fußballgott ‘footballing god’ this is true, for Wettergott ‘weather god’ it is false.

5 Experiments

5.1 Automatically extracted features

Our experimental set-up is as follows. We perform a 5-fold cross validation. Our data is initially randomized before we defined folds that are held constant across all experiments. For our experiments,
we convert the instances of B(oth) to instances of Y, i.e. affixoid uses. The seven instances that the annotators had left as unsure are excluded from the experiments. We use the automatically extractable features discussed in §4.1. As our classifier, we use SVM (Vapnik, 1995), as implemented by SVMlight (Joachims, 1998). Note that we experimented with the cost-factor, by which training errors on positive examples (i.e. examples of class Y) outweigh errors on negative examples (i.e. examples of class N). However, modifying it only helped in the lowest-performing configurations where we would otherwise perform exactly like the majority baseline. Table 6 reports results for the different features obtained with the default settings of SVMlight. The majority baseline we report represents a classifier that always predicts non-affixoid as this is the majority class for each suffixoid and the dataset as a whole.

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>Y (i.e. affixoid use)</th>
<th>N (i.e. non-affixoid use)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>P R F1</td>
<td>P R F1</td>
</tr>
<tr>
<td>all features</td>
<td>0.78</td>
<td>0.74 0.53 0.62</td>
<td>0.79 0.90 0.84</td>
</tr>
<tr>
<td>supersenses_bag</td>
<td>0.74</td>
<td>0.69 0.43 0.53</td>
<td>0.76 0.90 0.82</td>
</tr>
<tr>
<td>supersenses_diffs</td>
<td>0.66</td>
<td>0.29 0.13 0.18</td>
<td>0.67 0.93 0.78</td>
</tr>
<tr>
<td>psycholinguistic</td>
<td>0.72</td>
<td>0.69 0.30 0.41</td>
<td>0.72 0.93 0.81</td>
</tr>
<tr>
<td>frequency</td>
<td>0.71</td>
<td>0.67 0.34 0.44</td>
<td>0.73 0.91 0.81</td>
</tr>
<tr>
<td>emotion</td>
<td>0.67</td>
<td>0.67 0.08 0.14</td>
<td>0.67 0.97 0.79</td>
</tr>
<tr>
<td>pmi</td>
<td>0.66</td>
<td>0.02 0.16 0.04</td>
<td>0.66 0.98 0.79</td>
</tr>
<tr>
<td>majority</td>
<td>0.66</td>
<td>0.00 0.00 0.00</td>
<td>0.66 1.00 0.79</td>
</tr>
<tr>
<td>compositionality</td>
<td>0.66</td>
<td>0.00 0.00 0.00</td>
<td>0.66 1.00 0.79</td>
</tr>
<tr>
<td>polarity</td>
<td>0.66</td>
<td>0.00 0.00 0.00</td>
<td>0.66 1.00 0.81</td>
</tr>
</tbody>
</table>

Table 6: Results for classification experiments per feature group; setting: 5-fold CV

The bag of supersenses is the best feature group and notably better than paying attention only to differences between second component and complex word in terms of supersenses.

5.2 Automatic features in balanced setting

The difference in performance between the two classes that we found when using all data (cf. §5.1) might be due either to the class imbalance or come about because the majority class N is inherently easier to learn. In order to tease this apart, we here repeat the previous experiment using all features in a 5-fold cross-validation setting, however this time with a balanced dataset of 970 items. The dataset is constructed so that each fold contains the same number of instances per suffixoid and those instances are themselves balanced equally across the affixoid and non-affixoid classes. As Table 7 shows, with balanced data we can obtain much more balanced performance on the two classes. And despite the fact that we are using only 54.3% of the available data, the overall F1-score (0.73) is more or less tied with the result obtained on all data (0.74).

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>Y (i.e. affixoid use)</th>
<th>N (i.e. non-affixoid use)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>P R F1</td>
<td>P R F1</td>
</tr>
<tr>
<td>all features</td>
<td>0.73</td>
<td>0.71 0.78 0.75</td>
<td>0.76 0.67 0.71</td>
</tr>
</tbody>
</table>

Table 7: Results for classification experiment on balanced data using all features; setting: 5-fold CV

5.3 Feature ablation experiments

Table 8 shows an ablation experiment using all data. We remove one particular feature from the entire feature set at a time. The table shows that the performance decreases only marginally for most removed features, which indicates that most features encode information. The largest performance drop is caused by removing supersenses, which suggests that this is the feature with the most unique information.

Since the supersense feature is both the strongest individual feature (cf. Table 6) and also the feature encoding most unique information (cf. Table 8), we further examined this feature by listing the top 10 supersenses according to χ² ranking as shown in Table 9. The most predictive supersenses are predominantly those of the first component of the compound or those of the (complex) compound form itself. This first component is very helpful, for instance, for König. Its non-affixoid uses (e.g. Sarazenenkönig ‘king of the Saracens’) mostly co-occur with a first component denoting a Location or a nationality.

---

8These latter two constraints result in only 970 items being used rather than 984, if we simply used all 492 instances of class Y and added as many instances from class N.
Table 8: Ablation experiments where one feature is removed from the entire feature set.

Table 9: Top 10 supersense features according to \( \chi^2 \) ranking.

5.4 Cross-affixoid generalization

Using the automatic features, we experiment with a setting where we train on the instances of all but one affixoid candidate and then test on the instances of the remaining affixoid. For this experiment, we leave out the lexical Suffixoid feature. Table 10 shows the results.

Table 10: Results for cross-affixoid classification, training on six affixoids, testing on the remaining one; all features except suffixoid

The results show lower performance than in the cross-validation setting where we had instances of all affixoids in the train and the test folds. For \( \text{könig} \) the results drop down as low as the majority baseline in the present setting. However, the low result is very likely also heavily driven by the fact that the instances of \( \text{könig} \) make up close to 40% of our whole dataset. In other words, when testing on \( \text{könig} \), we use only 60% of the dataset to train on. On the other hand, we also see lower results on affixoids such as \( \text{bolzen} \) that are much less frequent than \( \text{könig} \) and for which we use more training instances in the present generalization setting than we did in the cross-validation setting in §5.1. This suggests that overall cross-affixoid generalization may be somewhat limited.

5.5 Gold-quality manual features

The second experiment we perform is a simple one: we use the manual features described in §4.3 and try to predict whether the complex form is an affixoid formation or not. Table 11 shows that all features except suffixoid

The results confirm that affixoids can be defined in terms of intensification and evaluation.
## Table 11: Results for classification on all data using manual features; setting: 5-fold cross-validation

<table>
<thead>
<tr>
<th></th>
<th>Acc.</th>
<th>Y</th>
<th>N</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>suffixoid intensity</td>
<td>0.96</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>suffixoid evaluation</td>
<td>0.93</td>
<td>0.92</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>word polarity</td>
<td>0.89</td>
<td>0.78</td>
<td>0.97</td>
<td>0.86</td>
</tr>
<tr>
<td>word intensity</td>
<td>0.89</td>
<td>0.76</td>
<td>0.98</td>
<td>0.86</td>
</tr>
<tr>
<td>word evaluation</td>
<td>0.89</td>
<td>0.78</td>
<td>0.97</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 11: Results for classification on all data using manual features; setting: 5-fold cross-validation

### 6 Error analysis

Table 12 shows confusion matrices per suffixoid candidate and for all data cumulatively for the best setting in our automatic experiments (cf. Table 6), in which all features were used in 5-fold cross validation. The breakdown per suffixoid candidate reveals that the overall performance is strongly influenced by König, which in particular contributes the bulk of the recall of true affixoid formations.

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bolzen</td>
<td>11</td>
<td>23</td>
<td>123</td>
<td>81</td>
<td>91</td>
<td>182</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bruder</td>
<td>1</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>17</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gott</td>
<td>28</td>
<td>91</td>
<td>127</td>
<td>82</td>
<td>91</td>
<td>174</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hai</td>
<td>1</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>17</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hengst</td>
<td>226</td>
<td>77</td>
<td>303</td>
<td>54</td>
<td>71</td>
<td>125</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>König</td>
<td>10</td>
<td>32</td>
<td>42</td>
<td>13</td>
<td>23</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaiser</td>
<td>319</td>
<td>288</td>
<td>607</td>
<td>121</td>
<td>409</td>
<td>1018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Confusion matrices on all data for the all features setting; gold: rows, predicted: columns

Table 13 below shows the confusion matrix for the experiment with a balanced data set reported in §5.2. As was to be expected based on the scores in Table 7, the performance across the different suffixoids is now much better. We can thus conclude that the difference in performance between the two classes that we found when using all data is largely owed to the class imbalance and that the distinction between affixoid formations and non-affixoid formations can in principle be learned for suffixoids other than König.

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bolzen</td>
<td>24</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>61</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bruder</td>
<td>12</td>
<td>8</td>
<td>20</td>
<td>5</td>
<td>231</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gott</td>
<td>14</td>
<td>14</td>
<td>28</td>
<td>7</td>
<td>77</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hai</td>
<td>24</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>61</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hengst</td>
<td>12</td>
<td>3</td>
<td>20</td>
<td>5</td>
<td>231</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>König</td>
<td>24</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>61</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaiser</td>
<td>20</td>
<td>10</td>
<td>30</td>
<td>10</td>
<td>400</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>11</td>
<td>19</td>
<td>30</td>
<td>10</td>
<td>400</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Confusion matrices on balanced data for the all features setting; gold: rows, predicted: columns

### 7 Conclusion

In this paper we studied German complex forms containing suffixoid candidates. To do so, we constructed a new data set with 1788 items distributed about 7 potential affixoid candidates, which we make available to the research community. In one set of experiments we validated the high correlation between evaluative and intensifying semantics and affixoid status that has always been assumed by theoretical work but not demonstrated. In another experiment, we tackled the task of classifying our complex forms as affixoids formations or regular compounds. Being able to do so successfully can be useful for sentiment analysis, as the affixoid uses are typically evaluative. The task is difficult, though, as we have little reliable information about the complex forms available due to their low frequency. Still, we achieved best results of 74% F1-score using a custom set of features, among which supersenses had the most impact.

In future work, we plan on pursuing lines of research that we needed to leave open here. First, we would like to cover more suffixoid candidates also extend this work to prefixoid formations. Second, once we have annotated data for further affixoid candidates in hand, we would like to explore if it is possible to identify morphemes that have affixoid uses from properties of their formations, including their frequency distribution, and distinguish them from morphemes that only participate in regular compounds. Since so far affixoids have been identified based on human introspection, we do not know how many of them actually exist in German or other languages.
References


Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws - a publicly available german-language resource for sentiment analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta, may. European Language Resources Association (ELRA).


