# A Corpus of Literal and Idiomatic Uses of German Infinitive-Verb Compounds

**Andrea Horbach°, Andrea Hensler⋆, Sabine Krome⋆, Jakob Prange°, Werner Scholze-Stubenrecht×, Diana Steffen°, Stefan Thater°, Christian Wellner°, Manfred Pinkal°**

°Department of Computational Linguistics, Saarland University, Saarbrücken, Germany,
{andrea, jprange, dsteffen, stth, chwell, pinkal}@coli.uni-saarland.de

⋆Redaktion Wahrig bei Brockhaus, Gütersloh, Germany,
{BMBF-Wahrig, Sabine.Krome}@bertelsmann.de

×Bibliographisches Institut GmbH Dudenverlag, Berlin, Germany,
Werner.Scholze-Stubenrecht@duden.de

**Abstract**

We present an annotation study on a representative dataset of literal and idiomatic uses of infinitive-verb compounds in German newspaper and journal texts. Infinitive-verb compounds form a challenge for writers of German, because spelling regulations are different for literal and idiomatic uses. Through the participation of expert lexicographers we were able to obtain a high-quality corpus resource which is offered as a testbed for automatic idiomaticity detection and coarse-grained word-sense disambiguation. We trained a classifier on the corpus which was able to distinguish literal and idiomatic uses with an accuracy of 85%.

**Keywords:** Corpus Annotation, Semantics, Idiom Detection

## 1. Introduction

This paper presents a study on German infinitive-verb compounds with an inflected head verb and an infinitive modifier. Specifically, we consider infinitive-verb compounds with the head verbs *bleiben* and *lassen* which allow for an idiomatic, i.e., conventionalized figurative interpretation in addition to a literal interpretation. Sentence (1) illustrates the lexicalized idiomatic sense "to have to repeat a year in school" of the compound *sitzenbleiben*, while (2) illustrates the semantically transparent use of the compound ("remain seated").

(1) Die Atlas-Werte bemessen sich unter anderem danach, wie gut Grundschüler lesen können, wie viele Schüler *sitzenbleiben*, wie viele Hochschulabsolventen eine Region zählt.

(2) Einmal schildert Grosz das Beispiel einer Frau namens Marissa Panigrosso, die in ihrem Büro im New Yorker World Trade Center nach dem Einschlag des ersten Flugzeugs am 11. September 2001 vom Schreibtisch aufsprang und durchs Treppenhaus auf die Straße flüchtete, während ihre liebsten Kolleginnen einfach im Büro *sitzen blieben*.

The distinction between literal and idiomatic uses is a prerequisite for correct spelling. According to the current German spelling rules[1], literal uses of infinitive-verb combinations must be written as two separate words (as in (2)), while idiomatic uses of compounds with *bleiben* and *lassen* as inflected head verbs can also be written in one word (as in (1)).

Our motivation for addressing this task is two-fold: the problem of idiom recognition is an interesting challenge from the point of view of computational semantics; due to the influence of the idiomaticity status on spelling, there is a societal and political motivation as well.

Spelling became a political issue in Germany through the German orthography reform, which was decreed in 1996 and caused a fierce controversy carried out in public as well as uncertainty even among professional writers about the correct spelling. As a consequence of this tumult caused by the reform, the Council for German Orthography (Rat für deutsche Rechtschreibung) was established in 2004. Its central responsibility is to observe and analyze whether the writing community follows the official orthographic regulations and to make recommendations for future changes. To provide evidence for this, the council has to continuously monitor the writing practice of professional as well as non-professional writers.

The work presented in our paper has been carried out in a project[2] which has the aim to provide corpus and NLP tools for the monitoring task of the council. A standard task in this context is the detection of instances that are relevant for monitoring how individual spelling rules are applied. Within this paper, we focus on verb-verb compounds. Their writing was affected by the spelling regulations and by subsequent modifications. Since correct spelling of verb-verb compounds with *bleiben* and *lassen* as head verbs depends on their context-specific semantic use, they are difficult for writers of German, and a challenge for automatic classification of relevant cases at the same time.

We report two relevant results: First, we have collected a corpus of 6,000 instances of 6 representative infinitive-verb compounds in German, double-annotated for idiomaticity by highly qualified experts for this task, i.e., the responsible editors of the two most important monolingual German dictionaries, Duden and Wahrig.[3] We obtained substantial inter-annotator agreement and at the same time were able to identify systematically difficult cases that could not be resolved even by the experts.

[1] `http://www.rechtschreibrat.com`

[2] "Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen", funded by BMBF (Project Number: 01UG1355B), `www.schreibgebrauch.de`

[3] Bibliographisches Institut GmbH Dudenverlag `http://www.duden.de` (Duden) and Redaktion Wahrig bei Brockhaus; Verlag F.A. Brockhaus / wissenmedia in der inmediaONE] GmbH (Wahrig)

Second, we trained a Naive Bayes classifier that uses various context features to classify instances of the verb compounds as either idiomatic or literal. We reached an average accuracy of $0.85$. The classifier will support lexicographers in the selection of relevant cases in monitoring the practice of German writers.

## 2. Corpus and Annotation

| Verb | Two words | One word |
|---|---|---|
| hängen+bleiben | 2840 | 2660 |
| liegen+bleiben | 3523 | 1803 |
| sitzen+bleiben | 2795 | 1325 |
| sitzen+lassen | 1994 | 450 |
| stehen+bleiben | 6932 | 4344 |
| stehen+lassen | 5361 | 1364 |

Table 1: Verb frequencies in the Wahrig corpus

### 2.1. Dataset

Our dataset is selected from the Wahrig corpus (Krome, 2010), a 3 billion word corpus of professional German writings that covers newspaper and magazine articles originating from between 1993 and 2013. Our dataset contains a total of 6,000 instances of six frequently used infinitive-verb compounds, which occur in literal as well as idiomatic uses and can accordingly be found in both orthographic variants. Table 1 shows the six compounds considered in this paper and gives the frequency of each orthographic variant in the Wahrig corpus.

For the 6 considered verbs, 1,000 instances of each compound, 500 per spelling, were randomly extracted across all years and publication media. (In the case of *sitzenlassen*, the Wahrig corpus contains only 450 relevant instances.) Note that according to German word order rules, the elements of the infinitive-verb compound often occur in inverted order and separated by different linguistic material, as in *Ihre liebsten Kolleginnen blieben einfach im Büro sitzen*. These cases obviously do not give rise to two orthographic variants. Therefore, we only considered cases where infinitive and head verb occur adjacent to each other and the head verb comes last (“rechte Satzklammer“).

### 2.2. Annotation process

Our annotators are two expert lexicographers, one each from Duden and Wahrig, with years of experience in the fine-grained analysis of word senses based on the inspection of usages of a lemma in real-world text. We did not specify explicit annotation guidelines, because we wanted to build on the expertise of the annotators. The annotators saw the sentence containing the verb and one sentence to the left and right as additional context and annotated each instance as either *literal* or *idiomatic*. They could use a question mark (*?*) to indicate cases where they were not sure about the label.

The annotation task was challenging for two reasons: First, there is no clear boundary between *literal* and figurative, or *idiomatic*, uses. There are borderline cases, where criteria do not apply, or they contradict each other. Second, we do not have two very clearly separated uses of the respective verbs, rather most of the verb groups have several literal and several figurative meanings, which we illustrate by means of the example of *stehen+bleiben*.

Sentences (3) and (4) represent two different literal uses: (3) talks about bowling pins that remain in an upright position, (4) about a person who has stopped walking (stands still).

(3) Zwei der neun Kegel blieben stehen.

(4) Er ist plötzlich stehen geblieben.

Sentences (5) to (7) indicate some idiomatic uses (out of a larger number): A person’s heart may stand still (5), people may stand still in their mental development (6), and you can claim that a statement cannot “remain standing”, i.e., remain uncontradicted, although in neither case are the spatial conditions for using the verb *stehen* / “stand” met (7).

(5) Ihm sei dabei das Herz stehen geblieben.

(6) Durch die Sucht ist er geistig stehen geblieben.

(7) Diese Aussage kann so nicht stehen bleiben.

### 2.3. Data analysis

| | L–L | I–I | ?–? | L/I–? | L–I | $\kappa$ |
|---|---|---|---|---|---|---|
| hängenbleiben | 96 | 362 | 2 | 40 | 13 | 0.79 |
| hängen bleiben | 144 | 306 | 3 | 47 | 9 | 0.81 |
| liegenbleiben | 163 | 273 | 4 | 60 | 21 | 0.77 |
| liegen bleiben | 260 | 189 | 4 | 47 | 19 | 0.82 |
| sitzenbleiben | 74 | 399 | 6 | 21 | 0 | 0.87 |
| sitzen bleiben | 227 | 214 | 10 | 49 | 7 | 0.82 |
| sitzenlassen | 1 | 429 | 0 | 20 | 0 | 0.25 |
| sitzen lassen | 43 | 408 | 3 | 46 | 5 | 0.66 |
| stehenbleiben | 201 | 233 | 6 | 60 | 33 | 0.78 |
| stehen bleiben | 243 | 191 | 1 | 65 | 37 | 0.75 |
| stehenlassen | 211 | 186 | 5 | 98 | 39 | 0.65 |
| stehen lassen | 217 | 179 | 2 | 102 | 48 | 0.63 |

Table 2: Results of our annotations: We report for each verb the number of cases where the two annotators agreed on literal meaning (L–L), on idiomatic meaning (I–I), how often they were both unsure (?–?) , the number of disagreements involving a question mark (L/I–?), disagreements between literal and idiomatic (L–I), and IAA values.

Table 2 shows an overview of the annotation results. We see in almost all cases high kappa values on the three-way classification, indicating *substantial* ($0.6 < \kappa < 0.8$) or *almost perfect agreement* ($\kappa > 0.8$). Low agreement values for *sitzenlassen* are explained by the extreme skewedness of the distribution in this sub-corpus.

Additionally, Table 3 shows the confusion matrix between the two annotators. Highlighted in red are cases of hard disagreement where one annotator decided for literal and the other for idiomatic meaning, and vice versa.

|   | L | I | ? |
|---|---|---|---|
| L | **1880** | 101 | 45 |
| I | 130 | **3369** | 239 |
| ? | 105 | 35 | 46 |

Table 3: Confusion matrix between our two expert annotators across all 6 compounds

**Adjudication.** After the annotation process had been finished, a subgroup of the authors, including the annotators, met in order to discuss the instances where they disagreed, including all cases where at least one annotator assigned a question mark. For time reasons, only parts of the data (236 out of 701 items) could be adjudicated. As a result of the discussion, we reached agreement on 79% of these items, classifying them as either literal or idiomatic. The main reason for the disagreement in these cases had been a different understanding of the scope of the literal use. The remaining 21% of items could not reliably be categorized as either literal or idiomatic.

Most of the instances undergoing adjudication can be grouped into a few larger groups. In a first subset of cases, uncertainty arises because it is unclear which criteria must be satisfied to make an instance a proper literal use of the verb group.

(8) Unter ihnen waren 10 Kinder, deren Schulbus den Fluss gerade noch überqueren konnte, aber auf einer eingebrochenen Rampe *hängenblieb*.
(Among them were 10 children whose school bus was able to cross the river but was *left hanging* at a broken ramp.)

In Example 8, the issue is whether the verb *hängen* ("to hang") requires a subject that gets stuck at a reference object in a vertical position. Another problem case was *hängen bleiben* regarding a football boot in a lawn. Most of these cases could be resolved (mostly towards literal use), but it turned out to be impossible to give a consistent definition of what "literal use" means even for a single lemma.

In a second frequent set of uncertain instances, uncertainty arises from the fact that the use of the verb group is literal and figurative at the same time. A typical example is Sentence 9. A local literal use *Fische bleiben im Netz hängen* is embedded in a more global metaphorical context. We decided to classify all instances of this kind as literal.

(9) Trotzdem sind viele Iraner der Meinung, dass nur kleine Fische im Netz der Drogenbekämpfung *hängen bleiben*.
(Nevertheless, many Iranians think that only the little fish *are taken up* ("are left hanging") in the net of the drug war.)

In addition to these uncertain cases, there are cases where classification is impossible or does not make sense at all. Some instances come with too little context information to make a decision (annotators were shown the sentences preceding and following the target sentence only). Some other instances are cases of semantic underspecification: the verb group is used literally and idiomatically at the same time, as *stehen lassen* in (10).

(10) Dem Hochzeitsunterhalter Robbie kommt der Frohsinn abhanden, als ihn die eigene Braut vor dem Traualtar *stehenlässt*.
(The wedding entertainer Robbie loses his own chance at happiness when his own bride *leaves him standing* before the altar.)

## 3. Experiments

This section presents results in automatically distinguishing literal and idiomatic uses of German infinitive-verb compounds. Previous approaches to detecting idiomatic expressions on the level of individual instances often relied on properties which are not applicable in our case, like the concepts of *canonical form* (Cook et al., 2007; Fazly et al., 2009) or *lexical cohesion* (Sporleder and Li, 2009). Conceptually, our work is more in the spirit of Birke and Sarkar (2006), who treated "literal" and "idiomatic" as two different senses of the target word and applied (unsupervised) word-sense disambiguation techniques. In contrast to this work, however, we use a simple supervised approach to distinguish literal from idiomatic uses.

### 3.1. Feature sets and classifier

The Wahrig corpus already comes with automatically annotated sentence and token boundaries, as well as POS information, so no additional linguistic preprocessing was needed. Inspired by standard approaches to supervised word sense disambiguation, we use the (lemmatized) words which occur within the same sentence as the target word as boolean features in our *basic* feature set. We consider only content words (nouns, verbs except for auxiliaries, adjectives, adverbs and prepositions) which occur at least three times.

Additionally, we tested several other groups of features. We use *local skip n-grams*, 1 to 6-grams in a window that spans 3 positions to the left and the right of the compound to be classified, skipping the compound itself. Additionally, we collect as features POS information of context words in the same window (*pos*) and the POS tag of the compound itself ($pos_0$) or the second part if it is written as two individual tokens. To cover syntactic information of an item, we use the subject and accusative object of the compound as assigned by the Zurich parser (Sennrich et al., 2009), as well as their part-of-speech tags (*syn*). All these features have been proposed by (Lee and Ng, 2002) for a word sense disambiguation task.

We also use selectional preference information (*sel*), counting how frequently the subject and accusative object head noun in a specific occurrence of the verb group occur as the subject or object, respectively, of the base verb (first component). For example, *Bild* (picture) occurs often as a subject of *hängen*, which we take as an indicator that an instance of *hängen bleiben* with *Bild* as a subject is used literally.

Finally, we add topical information (*topic*) for the news article in which an item occurs using manually annotated topic categories in the Wahrig corpus.

As our classifier, we use Naive Bayes (cf. Lee and Ng (2002)), from the Weka toolkit (Hall et al., 2009), as most of our features are independent categorical/nominal features.

### 3.2. Results

Table 4 shows the results for each of the six verbs. We report accuracy values using ten-fold cross-validation on the data on which both annotators agreed. As baseline we use a classifier that always assigns the majority class, and compare it to the simple bag-of-words feature set (*basic*) as well as the *full* dataset. The basic bag-of-words features are by far most effective. They consistently beat the majority baseline by a large margin of up to 30 percentage points, except for the extremely skewed data set for *sitzen+lassen*. Using the full feature set improves accuracy by another 2.5%.

To see the influence of individual groups of features, we performed an ablation test (see Figure 1): Disappointingly, selectional preference information has a slightly negative effect. A possible reason is that most base verbs can be used idiomatically as well. The POS features also do not help, probably because the part of speech of context verbs is not discriminative for our classification task. We removed those two features and retrained the classifier on the remaining feature set, which we call *selection*.

As can be expected, most of the time our *selection* feature sets perform best: adding these features gives us an average increase of 2.9% accuracy on top of the results of the basic model, so that we reach an average accuracy of 86.1% (+20.6% compared to the baseline). Strictly speaking, we can only claim the 85.4% reached by the *full* feature set, since we did not use a held-out development dataset for the feature selection step.

We also re-ran the experiments on the dataset that additionally contained those cases that were unclear at first, but could be adjudicated. The impact of the individual feature groups was very similar; however, the overall accuracy dropped by 1 to 3 percentage points. This is what one would expect, because the additional adjudicated data represent cases that are difficult to classify for a machine for the same reasons they are for a human.

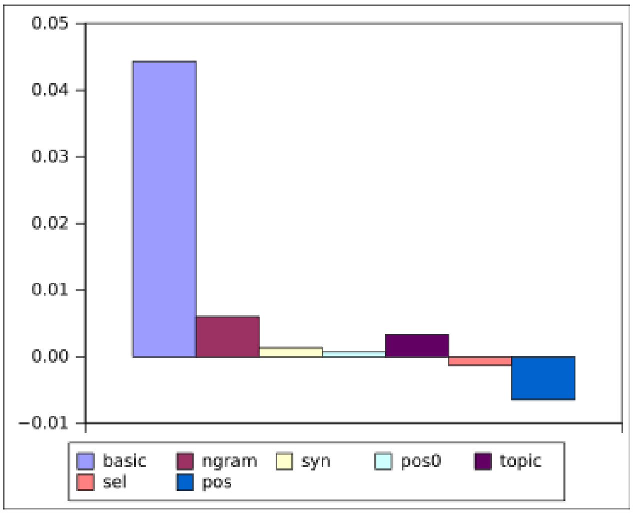

Figure 1: Ablation test for our features. y-values are the accuracy points lost by subtracting the respective feature group.

| | Baseline | basic | Full | Selection |
|---|---|---|---|---|
| hängen+bleiben | 0.736 | 0.826 | 0.822$^{**}$ | **0.836** |
| liegen+bleiben | 0.522 | 0.823$^{**}$ | **0.859** | 0.847$^{*}$ |
| sitzen+bleiben | 0.671 | 0.858$^{**}$ | 0.858 | **0.875**$^{*}$ |
| sitzen+lassen | **0.950** | 0.930 | 0.946$^{*}$ | 0.946 |
| stehen+bleiben | 0.512 | 0.728$^{**}$ | 0.805$^{**}$ | **0.812** |
| stehen+lassen | 0.540 | 0.826$^{**}$ | 0.832 | **0.847** |
| mean | 0.655 | 0.832 | 0.854 | **0.861** |

Table 4: Accuracy scores for the classification of literal and idiomatic uses. Results marked with $*$ are statistically significant with $p \leq 0.05$, those marked with $**$ significant with $p \leq 0.01$ compared to the next smaller value in the row (McNemar's test). *Baseline:* majority class guesser, *Full:* basic + n-grams + pos + syn + sel + topic, *Selection:* basic + n-grams + syn + topic

## 4. Application

In this section, we discuss how the classification results can be used for the monitoring task and related purposes. As an additional question about the usability of our classifiers in the future, we also investigate whether data from a certain range of years can be classified using only data from previous years as training.

### 4.1. Extracting relevant examples

The basic use case is the focused search for relevant examples. Our classifiers are able to distinguish between literal and idiomatic uses of infinitive-verb compounds with quite a high degree of precision. We assume that the classifier will be helpful for users who want to get lists of either literal or idiomatic uses for manual inspection: Some false positives may occur, but can be sorted out manually.

We can further increase precision at the cost of recall, by discarding instances to which the classifier assigns a confidence score below a certain threshold. We set different confidence thresholds, and evaluated only those instances whose confidence was above the threshold (the confidence values range between 0.5 and 1.0.).

Figure 2 shows the trade-off between confidence and the number of labeled data items: We can achieve a subsample of classified instances that contains 50% of the data and is more than 97% correct if we consider only those items about which the classifier has a confidence of 1.

### 4.2. Observing quantitative trends

Our automatic classification can also serve as a basis for the automatic detection of general trends in the acceptance of spelling rules and the development of spelling habits over the years. To demonstrate this, we first look at a frequency plot of the spelling variants for both literal and idiomatic uses based on our manually annotated gold standard data (see Fig. 3).

We see the effect of the spelling reform of 1996 (temporally

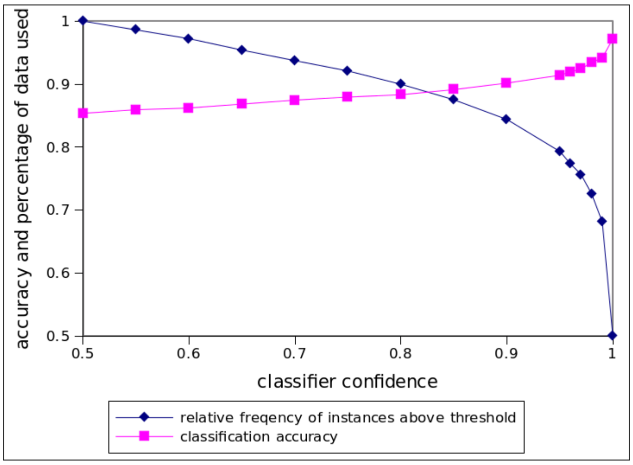

Figure 2: Trade-off between classification accuracy and the number of labeled data items with at least that accuracy.

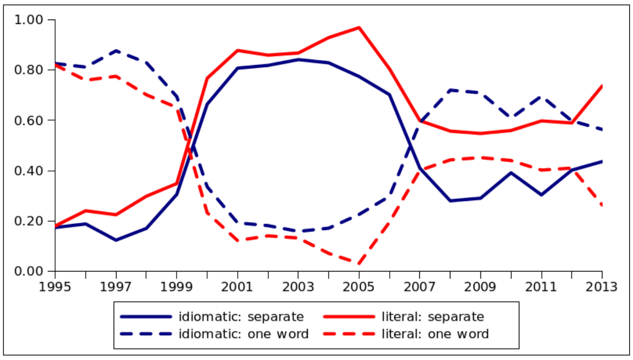

Figure 3: The relative frequency of separate writing among idiomatic and literal verb usages for individual years, computed on the gold standard data.

delayed by two or three years), which made separate writing obligatory for all verb-verb groups. In 2006, the regulations were modified for figurative or idiomatic uses, for which writing as one word was permitted (though not required). Accordingly, we see an increase of the frequency of variants written as one word for idiomatic uses. However, we also observe that literal uses start to be more frequently written as one word, though not to the same extent as idiomatic uses.

Fig. 4 shows the corresponding plot based on the annotation obtained from our automatic classifier. Although the curves are different, they show the same trends as those of Fig. 3 do: a sharp increase of separate writing after the 1996 reform, an increase of one-word writing for idiomatic uses after 2006, and a likewise significant, but lesser increase for literal uses.

### 4.3. Stability of the model across years

Our learned classifiers will be used to automatically label new corpus instances in coming years. It is therefore interesting to see how changes in language use over the years influence the classification accuracy. For example, context words are an important group of features in our classifier and they are sensitive to topics that might be more prevalent in some time periods. Therefore, we checked whether data from a certain range of years are labeled appropriately by a classifier that is trained on data from previous years, or whether a classifier has to be retrained on training data from the same range of years.

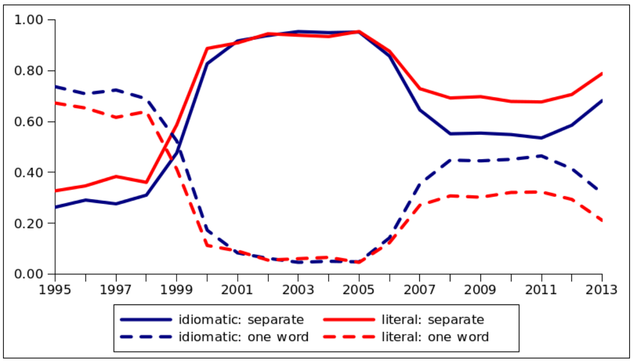

Figure 4: The relative frequency of separate writing among idiomatic and literal verb usages for individual years, computed on the automatically classified corpus.

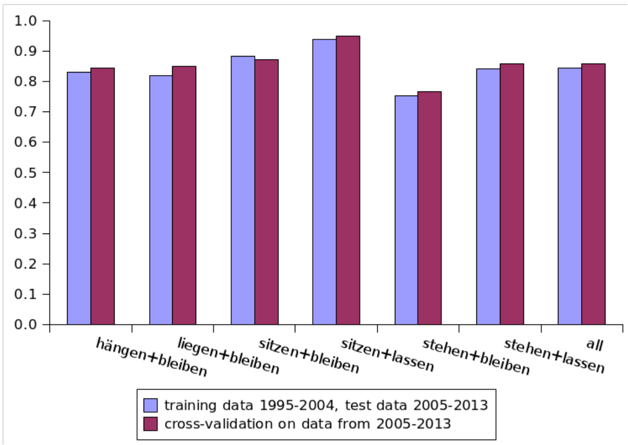

Figure 5: Classification accuracy on test data from 2005-2013 for a model trained on previous years (left column), or on the same range of years (right column).

For this purpose, we separated our annotated training data into two approximately equally sized subcorpora: those dating from up to 2004 and those published later. (We chose this date as it separates the data into two equal portions; it does not have a particular importance with respect to the history of spelling rules). We train classifiers on all items up to 2004 and test on the items after 2004 and cross-validate on the newer items for comparison. We can see that the difference between the two classifiers is minimal and of no practical importance.

## 5. Conclusions

In this paper, we present an annotation study on a representative dataset of literal and idiomatic uses of infinitive-verb compounds in German newspaper and journal texts. Through the participation of expert lexicographers we were able to obtain a high-quality corpus resource which is offered itself as a testbed for automatic idiomaticity detection and coarse-grained word-sense disambiguation. We trained a classifier on the corpus which was able to distinguish literal and idiomatic uses with an accuracy of 85%.

The classifier and other tools developed in the project have recently been handed over to the Council for German Orthography, where they will be used for the focused search for interesting spelling instances, and for the automatic detection of changes in spelling habits. The annotated corpus is made available[4] at Saarland University, Department of Computational Linguistics.

## 6. Acknowledgements

This work is part of the BMBF-funded project “Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen”.

## 7. References

Birke, J. and Sarkar, A. (2006). A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL’06)*, pages 329–336, Trento, Italy.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48. Association for Computational Linguistics.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Krome, S. (2010). Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie. Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des Wahrig Textkorpus digital. *Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive.*, pages 117 – 134.

Lee, Y. K. and Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 41–48. Association for Computational Linguistics.

Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for german. In *Proceedings of GSCL Conference*, Potsdam.

Sporleder, C. and Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics.

[4] `http://www.coli.uni-saarland.de/projects/schreibgebrauch/de/page.php?id=resources`