Classification of German verbs using nouns in argument positions and aspectual features

Michael Richter
Radboud University Nijmegen
mprrichter@t-online.de

Jürgen Hermes
University of Cologne
hermesj@uni-koeln.de

Abstract

This paper provides evidence that aspectual verb classes (Vendler, 1967) can be induced from nominal fillers in argument positions and aspectual features. We classified 35 German verbs in a supervised learning procedure using a support vector machine classifier and a classification into five aspectual classes (Richter and van Hout, 2015) as gold standard and observed excellent and substantial agreements.

1 Introduction

This study aims to empirically validate aspectual verb classes in German using large corpus data. Siegel (1997) and Siegel and McKeown (2000) induced the two aspectual classes *states* and *events* in the frame of a vector space model from corpora, however an induction of the complete Vendlerian typology has not yet been undertaken. We hypothesize that aspectual verb classes can be automatically induced from the classified nominal fillers in the argument position of verbs. Our hypothesis refers to the *Distributional Hypothesis* (Rubenstein and Goodenough, 1965; Schütze and Pedersen, 1995; Landauer and Dumais, 1997; Pantel, 2005) which says that semantically related linguistic elements appear in semantically related contexts. The present study in the framework of a vector space model is also driven by the *Statistical Semantics Hypothesis* (Weaver, 1955; Furnas et al., 1983; Turney and Pantel, 2010) which states that linguistic meaning can be derived from statistic linguistic patterns. In order to test our hypothesis, we took a test set of verbs from Schumacher (1986) and determined the nominal fillers and their classes in argument positions. That is, in subject, direct, indirect, and prepositional object positions by parsing a very large German corpus. As gold standard we used the aspect-based classification of Richter and van Hout (2015) into five classes which extends the typology of Vendler (1967), i.e. a*ccomplishments,* a*chievements, states* and *activities* by the class *accomplishments with an affected subject*.

This classification into five aspectual verb classes was derived by combining two user based classifications induced by cluster analyses from raters' judgments and associations with stimulus verbs and two usage based classifications induced from corpus data (Richter and van Hout, 2015). We took this classification as gold standard as we were interested in the correlation of the semantics of the nominal fillers in argument positions of verbs and the aspectual properties of verbs thereby following Klein (2009) who defines aspect as a grammatical category of verbs.

In the present study we represent verbs as vectors that consist of nouns in argument positions separated into areas according to their noun classes, which were induced by cluster analyses from similarity data. In addition, we added aspectual features as defined by Vendler (1967) to the vectors in order to compare the predictive power of the noun classes in argument positions against the predictive power of the aspectual features, respectively. The test set of verbs was classified in a supervised learning procedure using a support vector machine (SVM) classifier. In order to compare the results with aspectual verbs classes as gold standard with a gold standard-classification based on concrete semantic categories compatible with Schumacher's typology (1986) of German verbs, we trained the SVM classifier with a classification based on ten verb classes which comprises classes such as *verbs of consumption* and *verbs of handicraft working* (Richter and van Hout, 2015). This classification was induced

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

from the co-occurrence data bank (CCDB) of the Institut für Deutsche Sprache (IDS).[1]

2 Method

We classified 35 common German verbs used by Schumacher (1986), who defines seven lexical semantic macrofields and 30 subfields. We chose the verbs from all subfields, the only criterion being the representation of every subfield in order to cover the total semantic range of Schumacher's typology (1986). We checked the frequency of the verbs in the first one million sentences containing at least one of our selected verbs of the web based 880 million word SDEWAC corpus[2]. The verbs of our test set occurred in more than one million sentences with a mean frequency of approximately 30,000 occurrences per verb. 66 percent of the verbs was in the interval between 5,000 and 40,000 occurrences, the more frequent outliers being *müssen* 'to must' with 500,965 and *halten für* 'to take so./sth. for so./sth.' with 123,595 occurrences. We added five verbs; *hämmern* 'to hammer', *schneiden* 'to cut', *aufessen* 'to eat up', *laufen* 'to walk/to run', and *zersägen* 'to saw into pieces' since these verbs since a previous study (Richter and van Hout 2015) showed (i) that *laufen* 'to walk/to run' and *zersägen* 'to saw into pieces' are typical activity and accomplishment verbs respectively and (ii) that *aufessen* 'to eat up' is a typical accomplishment with an affected subject verb. *Schneiden* 'to cut' and *hämmern* 'to hammer' were ambiguous (Richter and van Hout, 2015), but we decided to classify in this study the former as accomplishment and the latter as a process verb.

In order to determine the verbs' arguments we parsed at most 30.000 sentences per verb using the Mate-Tools dependency-parser (Bohnet, 2010)[3]. The whole code we used for filtering and parsing the sentences, and aggregating the actants and aspectual features (see below) is available at GitHub.[4] The 35 verbs of our test set (Richter and van Hout, 2015) are represented as 139 dimensional vectors containing the 30 most frequent nouns in the verbs argument positions: subjects, direct objects, indirect objects and prepositional objects. The nouns were weighted by the TF-IDF measure and classified by cluster analyses carried out on a matrix with similarity values taken from the co-occurrence data bank (CCDB) of the Institut für Deutsche Sprache (IDS).[5] On the matrix of the similarity values, a cluster analysis with Ward's method and Euclidean distance was carried out. According to the *Bayesian Information Criterion* there are two optimal noun classes for all arguments. We interpreted the resulting noun classes using our intuition thereby applying the criterion of animacy (Croft, 2003; Aissen, 2003): The resulting two noun classes can be interpreted as denoting predominantly animate and inanimate things, respectively class $1_{[+\text{animate}]}$ for instance, contains nouns such as *Arzt* 'doctor' *Lehrkraft* 'teacher' and class $2_{[-\text{animate}]}$ contains nouns such as *Entwicklung* 'development', *Organisation* 'organization' and *Wahrnehmung* 'perception'. The verbs' vectors consist of areas for each argument type. There are four areas in total and each area is split into areas for each noun class as is depicted in (1):

$$\vec{v} = \begin{pmatrix} wn_1c_1 \\ wn_2c_1 \\ . \\ . \\ . \\ wn_n\ c_1 \\ wn_1\ c_2 \\ wn_2\ c_2 \\ . \\ . \\ . \\ wn_nv_n \end{pmatrix}$$

$(wn_i\ c_j\colon Weight\ of\ noun\ n_i\ in\ noun\ class\ \ c_j\ )$

Figure 1. *Dimensions of verb vectors*: *Weighted verbs in noun class areas.*

In addition, the vectors were completed by aspectual features that Vendler (1967) suggested in order to distinguish aspectual verb classes. The aspectual features indicate, for instance, whether the verbs occur in sentences with temporal specifications of duration or a limited time span with prepositions *in* and *for*, respectively, as in *he wrote the letter in an hour* versus *he wrote the letter for an hour*, whether the verbs can be embedded by matrix verbs such as *persuade* or whether they occur in imperative forms. In order to classify the 35 verbs we used a

[1] http://corpora.ids-mannheim.de/ccdb/. The similarity values were provided by Cyril Belica.
[2] The SdeWaC Corpus is available at the WaCky Corporadownload page at http://wacky.sslmit.unibo.it/doku.php?id=corpora
[3] See https://code.google.com/p/mate-tools/
[4] https://github.com/spinfo/verbclass
[5] The similarity values were provided by Cyril Belica.

SVM classifier with a non-linear kernel which achieved the best results.
We first trained the SVM using the classification of Richter and van Hout (2014) as a gold standard and tested it with a 10-fold cross-validation. The gold standard classification in detail:

1. accomplishments:
aufbauen auf 'to build on/to be based on', *herstellen* 'to produce', schneiden 'to cut', *zersägen* 'to saw into pieces', *verlängern* 'to extend', *mitteilen* 'to tell/to inform', *übermitteln* 'to communicate/to forward', *verhindern* 'to prevent', *abgrenzen* 'mark off/to define'
2. accomplishments with affected subject:
untersuchen 'to examine', *bedenken* 'to consider', *erörtern* 'to debate', *nachprüfen* 'to ascertain/to check', *aufessen* 'to eat up', *essen* 'to eat'
3. activities:
laufen 'to walk/to run', *eingehen auf* 'to respond to so./sth.', *hämmern* 'to hammer', *ansteigen* 'to increase'
4. achievements:
einschlafen 'to fall asleep', *vergehen* 'to go (by)/to pass/to diasappear', *übersehen* 'to overlook', *verlieren'to* loose', *anfangen* 'to begin', *abweichen* 'to deviate', *sich orientieren* an 'to be geared to', richten auf 'to direct towards/to focus'
5. states:
existieren 'to exist', *fehlen* 'to lack', *müssen* 'to must', *halten für* 'to take so./sth.for so./sth.', *folgen aus* 'to follow from', *angehören* 'to belong to', *übereinstimmen* 'to agree', *betreffen* 'to concern', *abweichen* 'to deviate' , *verhindern* 'to prevent'

The classification into classes of concrete lexical properties which we induced from the co-occurrence data bank (see above) is given below (the class labels are compatible with Schumacher's labels and are assigned using our linguistic intuitions; class 10 is incoherent and could not be labelled):

1. verbs of activities manipulating a substance (normally with a tool):
hämmern 'to hammer', *schneiden* 'to cut', *zersägen* 'to saw into pieces'
2. verbs of consumption:
aufessen 'to eat up', *essen* 'to eat'
3. verbs of difference, 'negative' processes, non-existence:
*müssen** 'to must', *einschlafen** 'to fall asleep' , *vergehen* 'to go (by)/ to pass/to disappear', *übersehen* 'to overlook', *fehlen* 'to lack', *verlieren* 'to loose', *verhindern* 'to prevent', *abgrenzen* 'to mark off/ to define', *abweichen* 'to deviate
4. verbs of transfer (of information):
mitteilen 'to inform', *übermitteln* 'to communicate/to forward'
5. verbs of examination (by mental activity):
nachprüfen 'to ascertain/to check', *erörtern* 'to debate', *untersuchen* 'to examine'
6. verbs of production:
aufbauen auf 'to build on'/acc to be based on', *herstellen* 'to produce'
7. verbs of beginning and rising processes:
anfangen 'to begin', *ansteigen* 'to rise/ to increase'
8. verbs of discussion and consideration:
betreffen 'to concern', *bedenken* 'to consider', *eingehen auf* 'to respond to so./sth.', *halten für* 'to take, *richten auf* 'to direct towards', *orientieren an* 'to be geared to'
9. verbs of membership and agreement:
angehören 'to belong to', *übereinstimmen mit* 'to agree with'
10. *folgen aus* 'to follow from', *laufen* 'to walk/to run', *existieren* 'to exist', *verlängern* 'to extend'

2.1 Results

In order to evaluate the consistency of the comparisons of the classifications against the gold standards we calculated both accuracy and Cohen's kappa. The latter measure considers the number of classes which differ in the two gold standards and, in addition, gives the significance levels.
Taking the classification with five aspectual verbs classes as gold standard the subject feature clearly outperforms the remaining features with .857 accuracy (which means that 30 of 35 verbs were classified correctly) and κ = .812. Kappa values above .61 are characterized as substantial, above .81 as almost perfect agreement and therefore highly significant. The combinations subject-direct object-prepositional object-aspectual features and subject-direct object-aspectual features yield .828 accuracy, κ = .775 and κ = .773, respectively. The combinations subject-prepositional object-aspectual features, subject-direct object-prepositional object and subject- aspectual features yield .8 accuracy each with κ = .741, κ = .739 and κ = .71, respectively.

In contrast the remaining features, including the aspectual feature which yields .514 % accuracy, with κ = .317 (fair agreement), perform poorly.
Taking the classification according to concrete semantic properties into ten classes as the gold standard we observed that the hierarchy remains almost the same, the subject feature outperforms the remaining features. However, the accuracy is considerably lower compared to the classification with 5 aspectual verb classes. The subject achieves .657 accuracy, κ = .573. The combinations subject-direct object-aspectual features and subject-direct object-prepositional object yield .628 accuracy with κ = .458, followed by the combinations subject-direct object and subject-aspectual features with .6 accuracy each and κ = .495. These combinations exhibit a moderate agreement. Again, the aspectual feature performs poorly with .428 accuracy, κ = .266 which is a fair agreement. In figure 2 the accuracy of the argument and aspectual features for the comparisons against both gold standard classifications are given.

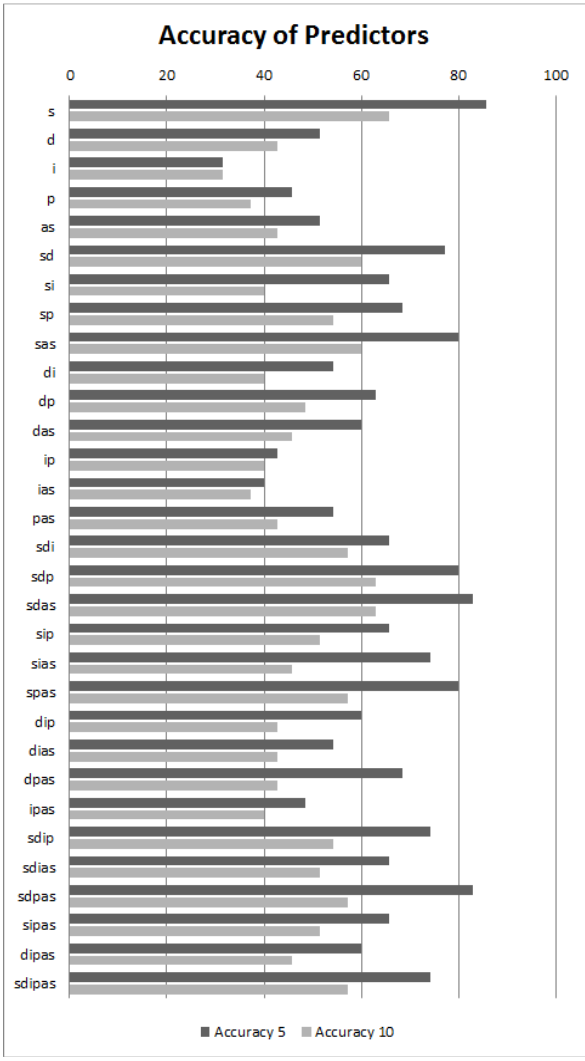

Figure 2. *Accuracy of the argument and aspectual features using five aspectual verb classes vs. ten classes with concrete lexical properties as gold standard.*

Note: s: subject, d: direct object, i: indirect object, p: prepositional object, as: aspectual features, and combinations of predictors, for instance, das: direct object and aspect, sp: subject and prepositional object.

3 Conclusion

The study provides evidence for the hypothesis that aspectual verb classes can be induced from classified nominal fillers in argument positions. For the five aspectual verb classes used as the gold standard (Richter and van Hout, 2014) it turned out that noun classes in subject positions have the highest predictive power compared to the nouns in the remaining argument positions and the aspectual features derived from Vendler (1967). This result is surprising since the Vendlerian aspectual categories were formulated in order to distinguish aspectual classes. Future research should explore a comparison of the predictive power of nominal and aspectual features.
Using a classification into concrete lexical fields as the gold standard of the predictive values we observed a considerable decrease in the predictive values indicated by the lower kappa values. We explain this result by the difference in information provided by the argument structures of the verbs in the 5 class-gold standard classification in contrast to the information provided by co-occurrences that is, lexical information of any type in the context of verb in the 10-class gold standard classification.
The results of this study show that: 1. Aspectual verb classes can be empirically validated, 2. Classified nouns in subject argument positions are reliable predictors of aspectual verb classes, i.e. the meaning of nouns in combination with their noun classes correlates with aspectual parts of the verbal meaning. In order to confirm these results further research with an extended test set of verbs is needed.

Acknowledgments

Roeland van Hout suggested to evaluate the classifications with Cohen‘s Kappa and was very helpful in the calculations of the κ-values.

References

Judith Aissen 2003. Differential Object Marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21: 435 – 483.

Bernd Bohnet 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

William Croft 2003. *Typology and Universals*. Cambridge University Press (2nd edition).

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais 1983. Statistical semantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal*, 62 (6): 1753 – 1806.

Wolfgang Klein 2009. How time is encoded. In W. Klein and P. Li (eds.), *The expression of time*, 39 – 82. Berlin: Mouton de Gruyter.

Thomas K. Landauer and Susan T. Dumais 1997. A solution to Plato's problem: The Latent Semanctic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104: 211 – 140.

Patrick Pantel 2005. Inducing ontological co-occurrence vectors. In *Proceedings of Association for Computational Linguistics* (ACL-05): 125 – 132.

Michael Richter and Roeland van Hout 2015. A classification of German verbs using empirical language data and concepts of Vendler and Dowty. To appear 2015 in *Sprache und Datenverarbeitung – International Journal for Language Data Processing.*

Herbert Rubenstein and John B. Goodenough 1965. Contextual correlates of synonomy. In *Communications of the ACM,* Vol 8 (10): 627 – 633.

Helmut Schumacher 1986. *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben.* Berlin & New York: De Gruyter.

Hinrich Schütze and Jan Pedersen 1993. A vector model for syntagmatic and paradigmatic relatedness. In *Making Sense of Words:* 104 – 113. Ninth Annual Conference of the UW Centre for the New OED and Text Research, Oxford.

Peter D. Turney and Patrick Pantel 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37: 141 – 188.

Zeno Vendler 1967. *Linguistics in Philosophy*. Ithaka /New York: Cornell University Press.

Warren Weaver 1955. Translation. In W. N. Locke, D.A. Booth, *Machine Translation of Languages*: 15 – 23. Cambridge Mass.: MIT Press.