Distributional Analysis of Function Words

Daniel Hole and Sebastian Padó
University of Stuttgart, Germany
{daniel.hole@ling, pado@ims}.uni-stuttgart.de

Introduction. Theoretical linguists observe with envy the way in which distributional semantics in computational linguistics renders research viable whose foundations were postulated by clear-sighted structuralists (Firth, 1957; Harris, 1954). Their interest diminishes upon seeing computational linguistics dealing mainly with parts of speech dominated by content words (nouns, verbs, adjectives), whereas theoretical linguists firmly believe that function words and morphosyntax define the interesting backbone of natural languages. The only part of speech containing a substantial number of function words that has received attention in the computational lexical semantics area is prepositions (Schneider et al., 2018).

This paper is a first attempt at reconciling the advanced tools of computation in distributional semantics with the function word emphasis of formal linguistics. We consider a multiply polysemous function word, the German reflexive pronoun *sich*, and investigate in which ways natural subclasses of this word which are known from the theoretical and typological literature (Kemmer, 1993, cf. Table 1) map onto recent models from distributional semantics. Due to the differences between lexical and functional polysemy, our preliminary results are different from those of studies of content word polysemy in distributional semantics (e.g., Boleda et al., 2012). We submit that our results open a window onto patterns of polysemy that may, in the long run, turn out at least as interesting and relevant to the computational study of natural languages as content words. What we find in our pilot is that some traditional subclasses of *sich* not only map neatly onto clusters produced by distributional methods, but that others which are predicted by theory to belong to constructional metaclasses with a wider distribution pervade the whole clustering space. What is more, the distribution of causative-transitive vis-à-vis anticausative verb types and of other verb classes partly reproduces the semantic map of the middle domain as first envisaged by Kemmer (1993) on a typological database. We take these results to be promising for more and in-depth studies of function morphemes in distributional semantics.

Distributional analysis and Data. Distributional analysis is probably the dominant paradigm for semantic analysis in computational linguistics. Building on the distributional hypothesis, *"you shall know a word by the company it keeps"* (Firth, 1957), they typically represent words as high-dimensional vectors representing the words' contexts and interpret vector similarity as semantic relatedness (Turney and Pantel, 2010). Virtually all work in this area has concentrated on *content* words (most common nouns, verbs and adjectives), following the intuition that these word classes refer to categories whose properties and relational structure can be learned profitably from distributional analysis (Cimiano et al., 2005).

In this paper, we focus instead on a function word form, namely the German reflexive pronoun *sich*. Traditionally, the context of function words was considered to be too general to be amenable to distributional analysis. The situation has changed with a generation of recently proposed

Category / Example	predict able	all persons	stress- able	+*lassen*	dispos- ition
1. INHERENT REFLEXIVES: *Paul schämte sich*/'Paul felt ashamed'	+	+	-	-	+/-
2. ANTI-CAUSATIVES: *Die Erde dreht sich*/'The earth revolves'	+	-	-	-	+/-
3. CHANGE IN POSTURE: *Paul setzte sich hin*/'Paul sat down'	+	+	-	-	-
4. TYPICALLY SELF-DIRECTED: *Paul kämmte sich*/'Paul combed his hair'	-	+	-	-	-
5. TYPICALLY OTHER-DIRECTED: *Paul erschoss sich*/'Paul shot himself'	-	+	+	-	-
6. DISPOSITIONAL MIDDLE: *Die Dose lässt sich leicht öffnen*/'The can opens easily'	+	+	-	+	+
7. EPISODIC MIDDLE: *Paul lässt sich beraten*/'Paul gets advice'	+	+	-	+	-
8. RECIPROCALS: *Die Geraden schneiden sich im Unendlichen*/ 'The lines intersect in the infinite'	-	-	+/-	-	+/-

Table 1: Salient uses of *sich*, inspired by Kemmer (1991).

distributional models that learn so-called *contextualized embeddings*. These models concurrently learn (a) general vectors for word types and (b) specialized vectors for word tokens in their context. This division of labor circumvents the generality problem: even if the representation of the word type *sich* is too general to be useful, the ability of the model to learn how the meaning of each *sich* token arises from a combination of basic word meaning and context.

The specific embedding model we use is BERT (Devlin et al., 2019), a so-called transformer architecture which captures relations among words in an unsupervised fashion with the help of an attention mechanism (Vaswani et al., 2017). In concrete terms, we use the pretrained 'BERT multilingual base' model which provides 768-dimensional contextualized embeddings for all input tokens. To visualize these vectors, we perform principal components analysis, a standard dimensionality reduction method, to represent instances of *sich* on a two-dimensional plane.

The basis of our analysis is the 700M token SdeWAC web corpus (Faaß and Eckart, 2013). We select the first 335 out of more than 5.5 million instances of *sich* for manual annotation by one of the authors with the eight classes as defined above. We experiment with two conditions of presenting the tokens in context to BERT: once we present them in their local *phrasal* context, as approximated by punctuation, and once in their complete *sentential* context. For visualization, we reduce the embeddings to two dimensions with principal components analysis.[1]

Findings. The two-dimensional instance representations for *phrasal* contexts are shown in Figure 1. In our estimation, the overall picture is promising: even though the classes are not completely separated, clear tendencies are visible.

- Inherently reflexive verbs (class 1) are interspersed through all event types and do not form a cluster of their own, as could be expected given their predictable nature. We therefore also show a figure with class 1 removed.

- Typically other-directed reflexive events like 'shooting oneself' and typically self-directed reflexive events like 'defending oneself' or 'combing' (classes 4, 5) form rather compact neighboring categories in the lower right sector.

[1]While three-dimensional representation would be possible, presentation on a page requires a two-dimensional visualization in the end.

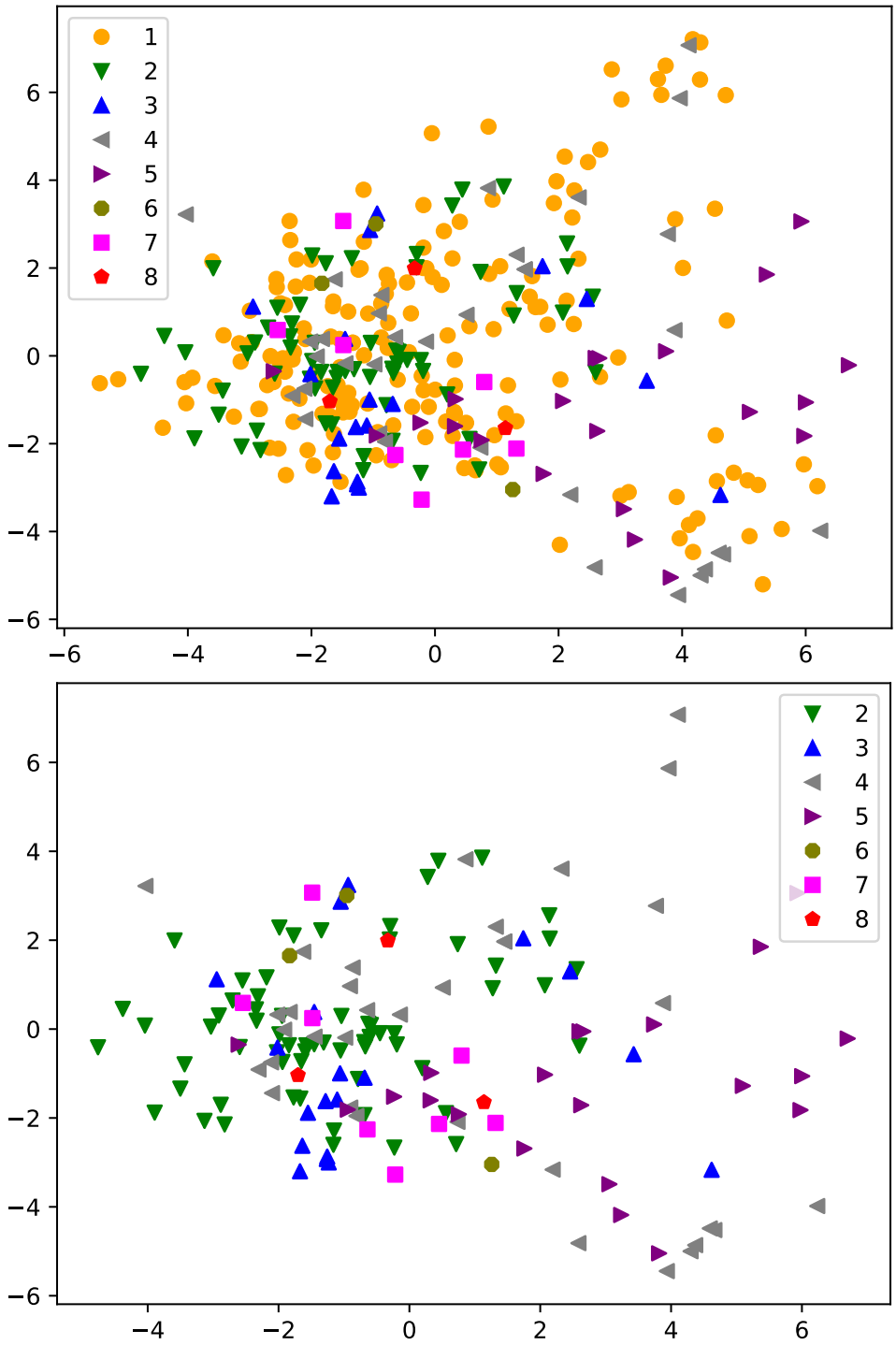

Figure 1: Distributional representations of *sich* instances based on phrasal contexts. All classes (above), without inherent reflexives (below). Class labels according to Table 1.

- The sectors on the right generally assemble agentive causative verb uses, whereas sectors on the left assemble anticausative verb uses like 'diminishing' or 'revolving' (class 2), all of which involve use of *sich* in German. Hence the bow from left to bottom right forms a path of growing agentivity, with traditional middle constructions (classes 3, 6, 7) literally occupying the middle of the plot.

- Some of the classes show a 'core' surrounded by outlier clouds. For the change-of-posture verbs (class 3), the noticeable string of outliers to the right is formed by non-literal uses (non-physical motion, e.g. *sich aus dem Verderben erheben* 'to rise from doom', *sich auf*

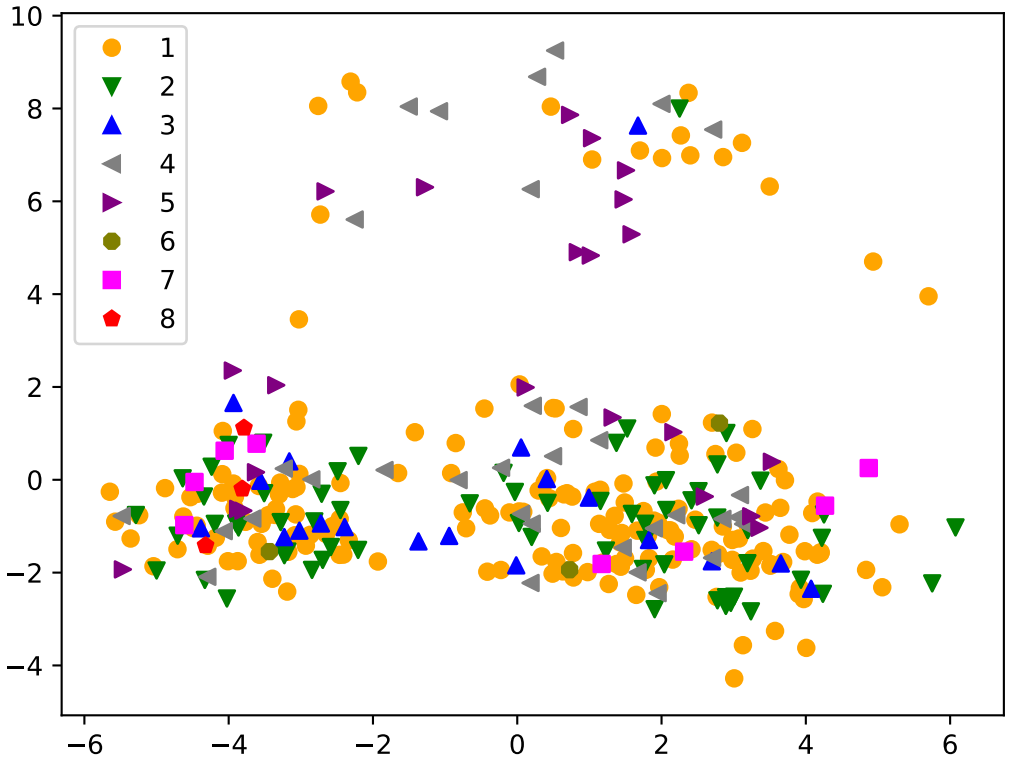

Figure 2: Distributional representations of *sich* instances based on sentential contexts.

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Class 2	0.98						
Class 3	0.96	0.95					
Class 4	0.98	0.95	0.95				
Class 5	0.93	0.89	0.91	0.93			
Class 6	0.91	0.91	0.88	0.88	0.83		
Class 7	0.94	0.92	0.93	0.92	0.89	0.90	
Class 8	0.89	0.86	0.88	0.88	0.84	0.83	0.86

Table 2: Inter-class similarities (phrasal context)

die Rechtsgrundlage stützen 'to rest on the legal foundation').

- The seemingly inhomogeneous behavior of the self-directed verbs (class 4) can be explained in terms of the distinction between PP-*sich* and DP-*sich* (Gast and Haas, 2008): The middle 'core' of class 4 consists of the DP cases, e.g. *sich unterziehen* 'to undergo'. In contrast, the cloud on the lower right is made up of PP cases like *bei sich tragen* 'to carry'. The latter are clearly more causative, in line with the 'causation' gradient described above. Finally, the outliers in the upper right sector are non-literal instances.

In contrast, Figure 2 shows the instance embeddings for *sentential* contexts. Here, the overall separation of instances in two horizontally separated clusters overshadows any separation by class label. We interpret this difference as an indication that a phrasal context (of on average 12 tokens) is sufficient to disambiguate *sich*, while in a full sentential context (of on average 77 tokens) the meaning of *sich* is overwhelmed by the meaning of the surrounding content words, as in traditional distributional investigations.

Finally, we present pairwise inter-class similarities for the phrasal context condition, as shown

in Table 2. These numbers are Cosine similarities, computed between class centroids in the original 768-dimensional embedding space, not the two-dimensional visualization space. Leaving again aside class 1 with its relatively uniform distribution, we observe that Classes 2, 3, 4, and 7 are relatively close to each other, as expected given the feature representations in Table 1, where many of the classes differ in only one feature. Class 5 is still relatively similar to class 4, which does not only fall out of the feature values, but also out of the fact that there is a gradient or grey area between classes 4 and 5. Classes 6 and 8 are both dissimilar to 5 and dissimilar to one another, but are otherwise too infrequent to draw strong conclusions.

References

Boleda, G., S. Schulte im Walde, and T. Badia (2012). Modeling regular polysemy: A study on the semantic classification of Catalan adjectives. *Computational Linguistics 38*(3), 575–616.

Cimiano, P., A. Hotho, and S. Staab (2005). Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research 24*, 305–339.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, Minneapolis, pp. 4171–4186.

Faaß, G. and K. Eckart (2013). SdeWaC – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, Volume 8105 of *LNCS*, pp. 61–68. Springer.

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. Oxford University Press.

Gast, V. and F. Haas (2008). On reciprocal and reflexive uses of anaphors in german and other european languages. In E. König and V. Gast (Eds.), *Reciprocals and reflexives: Theoretical and typological explorations*, pp. 307–346. Mouton de Gruyter.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2–3), 146–162.

Kemmer, S. (1993). *The Middle Voice*, Volume 23 of *Typological Studies in Language*. Amsterdam and Philadelphia: John Benjamins.

Schneider, N., J. D. Hwang, V. Srikumar, J. Prange, A. Blodgett, S. R. Moeller, A. Stern, A. Bitan, and O. Abend (2018). Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of ACL*, Melbourne, Australia, pp. 185–196.

Turney, P. D. and P. Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research 37*(1), 141–188.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Proceedings of NeurIPS*, Long Beach, CA, pp. 5998–6008.