Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5962–5967
Marseille, 11–16 May 2020

SeCoDa: Sense Complexity Dataset

David Strohmaier, Sian Gooding, Shiva Taslimipoor, Ekaterina Kochmar
University of Cambridge, Department of Computer Science and Technology, ALTA Institute
Cambridge, CB3 0FD
{ds858, shg36, st797, ek358}@cam.ac.uk

Abstract

The Sense Complexity Dataset (SeCoDa) provides a corpus that is annotated jointly for complexity and word senses. It thus provides a valuable resource for both word sense disambiguation and the task of complex word identification. The intention is that this dataset will be used to identify complexity at the level of word senses rather than word tokens. For word sense annotation, SeCoDa uses a hierarchical scheme that is based on information available in the Cambridge Advanced Learner's Dictionary. This way, we can offer more coarse-grained senses than directly available in WordNet.

Keywords: word sense disambiguation, lexical complexity, dataset

1. Introduction

Both word sense disambiguation (WSD) and complex word identification (CWI) are long established tasks in natural language processing (Navigli, 2009; Iacobacci et al., 2016), with CWI focusing on the detection of words that might be considered complex by readers and therefore in need of simplification (Shardlow, 2013). It has been shown that a number of NLP tasks benefit from WSD, however so far CWI and WSD have been kept distinct. The complexity differences between senses as opposed to the complexity of word tokens is yet to be specifically investigated or detected. The **Se**nse **Co**mplexity **Da**taset (**SeCoDa**) fills this gap by providing both complexity and sense information for word tokens. We release this dataset to open the door for a new combined task of *complex sense detection*.[1]

The SeCoDa also makes a contribution to word sense disambiguation research by moving beyond WordNet senses. Instead of WordNet, it uses more coarse-grained sense information found in the Cambridge Advanced Learner's Dictionary (CALD).[2]

2. Motivation

Our dataset is the first to combine word senses with complexity information. We argue that the same word with multiple senses may differ in its complexity when used in each of these senses. For example, the sense that the word *driver* takes in *car driver* may be deemed less complex than when it is used in the computer science-related sense. In everyday contexts, explaining what a driver in computer science does is considerably more difficult than what a car driver does. Examples from the dataset of Yimam et al. (2017) annotated with complexity scores support this idea. For instance, consider the following sentence:

> *Successive **waves** of bank sector clean-ups have failed to convince investors*

Waves in this context is used in the sense of "a larger than usual number of events of a similar, often bad, type, happening within the same period". This sense is less frequent than "a raised line of water that moves across the surface of an area of water, especially the sea". As a result, *waves* in this context is marked as complex in the dataset of Yimam et al. (2017), and is considerably more complex for the readers according to CALD.[3]

Existing work in complexity detection does not sufficiently address complexity differences at the level of word senses. None of the work on CWI has addressed assessing word complexity from the perspective of different word senses, despite noticeable differences in complexity levels the same word might take in different contexts. A shared task on Complex Word Identification was organized in 2018 (Yimam et al., 2018) on the basis of the dataset from Yimam et al. (2017). The winning system for the binary complexity task by Gooding and Kochmar (2018) considered word sense superficially, by incorporating the number of synsets available for the target word in WordNet.

When performing an error analysis on this system it was clear that context was crucial when determining word complexity. This led to a new state-of-the-art CWI system, SEQ, by Gooding and Kochmar (2019), where the left and right context of the target word is included using a bi-directional recurrent neural architecture. However, despite this system considering the sentential context of the target word, words are represented using GloVe embeddings (Pennington et al., 2014), thereby omitting fine-grained sense information. Table 1 illustrates an example where the word *capital* has alternate binary (BIN) annotations depending on the sense. The SEQ system, whilst outputting a slightly higher complexity probability in the first case, would still mark both of these words as not complex, because it uses 0.5 as the cut-off threshold and both occurrences have complexity scores below this value. Our data is intended to help develop a system that is able to recognise that these words have distinct senses when considering the corresponding complexity.

Our work fills this gap in the available resources by provid-

[1] The dataset can be found at `https://github.com/dstrohmaier/SeCoDa`.

[2] `https://dictionary.cambridge.org/dictionary/english/`

[3] See the Common European Framework of Reference for Languages (Council of Europe, 2011) levels for the entry `https://dictionary.cambridge.org/`

Sentence	BIN	SEQ
*The extra **capital** will have to be raised by the banks* [...]	1	0.43
*With that, it was back to the sprawling U.S. air base outside the **capital*** [...]	0	0.39

Table 1: Binary annotations for different senses of *capital*

ing annotation of sense and complexity for the same tokens.

Most existing datasets for word senses, such as the widely used SemCor and SensEval sets (Fellbaum, 1998; Edmonds and Cotton, 2001; Litkowski, 2004), are based on WordNet. Because WordNet is not solely focused on automatic WSD, there are multiple problems with its use for this task. One problem that has already received attention in the literature (Ide and Wilks, 2007) is the fine-grained nature of the distinctions drawn by WordNet. For an example, consider the five senses of *understand* WordNet provides (see Table 2).

Synset Name	Gloss
understand.v.01	know and comprehend the nature or meaning of
understand.v.02	perceive (an idea or situation) mentally
understand.v.03	make sense of a language
understand.v.04	believe to be the case
sympathize.v.02	be understanding of

Table 2: Available senses for *understand* in WordNet

The sense of *understand* as perceiving an idea mentally is close to knowing and comprehending the nature or meaning of something: the WordNet example sentences "I *understand* what she means" (for understand.v.01) and "I don't *understand* the idea" (understand.v.02), hardly help to distinguish these two synsets.

Even with an unusually large amount of context, prying those five senses apart might not be possible or even desirable. These fine distinctions could at best be drawn at a level where situational pragmatic considerations are taken into account, while automatic sense disambiguation might aim at a coarser level of meaning (Levinson, 1995; Levinson, 2000).

Our dataset addresses this problem by providing senses at two levels of granularity, both of which are coarser than WordNet synsets. We use the Cambridge Advanced Learner's Dictionary to achieve this goal. We note that using a dictionary to this end is not unusual: previously, the New Oxford American Dictionary has been put to similar use (Yuan et al., 2016). Another strategy used in the past involved trying to group WordNet senses to increase coarseness of granularity (Mihalcea and Moldovan, 2001; Agirre and Lacalle, 2003; Peters et al., 1998; McCarthy, 2006; Palmer et al., 2007). One advantage of CALD we do not exploit in the present dataset, but might be of use in the future is that this dictionary, provides CEFR information about word sense complexity for a subset of its entries.

dictionary/english/wave

3. Data Selection and Annotation Process

The dataset that we release with this paper contains both sense information and word complexity ratings. The annotation was undertaken in two separate steps and by two different sets of annotators. In particular, the complexity levels were assigned to words before the dataset was filtered for tokens that have sufficiently many senses for the purpose of word sense disambiguation.

3.1. Complexity Annotation

In this work, we re-annotate the dataset of Yimam et al. (2017) with word senses. The original dataset contains 30147 words annotated as complex or simple in context. The contexts in this dataset come from three different sources: professionally written NEWS, WIKINEWS written by amateurs, and WIKIPEDIA articles. All contexts were annotated by 20 annotators (including 10 native and 10 non-native English speakers) using the Amazon Mechanical Turk crowdsourcing platform. The annotators were asked to select the words in context that they find complex while reading. Text was presented to the annotators until 10 native and 10 non-native readers submitted their judgements, and the resulting word complexity annotation follows two settings. In the *binary* setting, a word is labelled with 1 if any of the 20 annotators marked it as complex, and it is labelled with 0 if none of the annotators selected it. In the *probabilistic* setting, words receive a complexity score that reflects the proportion of annotators, out of 20, that annotated it as complex. For instance, *ahead* used in the sense of "having more points, votes, etc. than someone else in a competition, election, etc." received a complexity score of 0.1 in this dataset when annotated in:

> *Team Germany won it with 128 points, 35 points lead **ahead** of Team New Zealand*

This means that 2 out of 20 annotators marked this word as complex. In contrast, *ahead* being used in the sense of "in or into the future" is not selected as a complex word in any of its contexts of use in this dataset, including:

> *There will be difficult days **ahead***

Notably, CALD also assigns different complexity levels to these two senses.[4]

3.2. Sense Annotation

The sense information is drawn from the Cambridge Advanced Learner's Dictionary (CALD). The complexity levels assigned to some of the entries in this dictionary are linked to the Common European Framework of Reference for Languages (CEFR) scheme (Council of Europe, 2011), which assigns levels A1-A2 to beginner, B1-B2 to intermediate, and C1-C2 to advanced learners of English.

The dictionary, including the full definitions, is available online. On the basis of our dataset the complexity of the disambiguated word senses can be investigated and the

[4] See https://dictionary.cambridge.org/dictionary/english/ahead

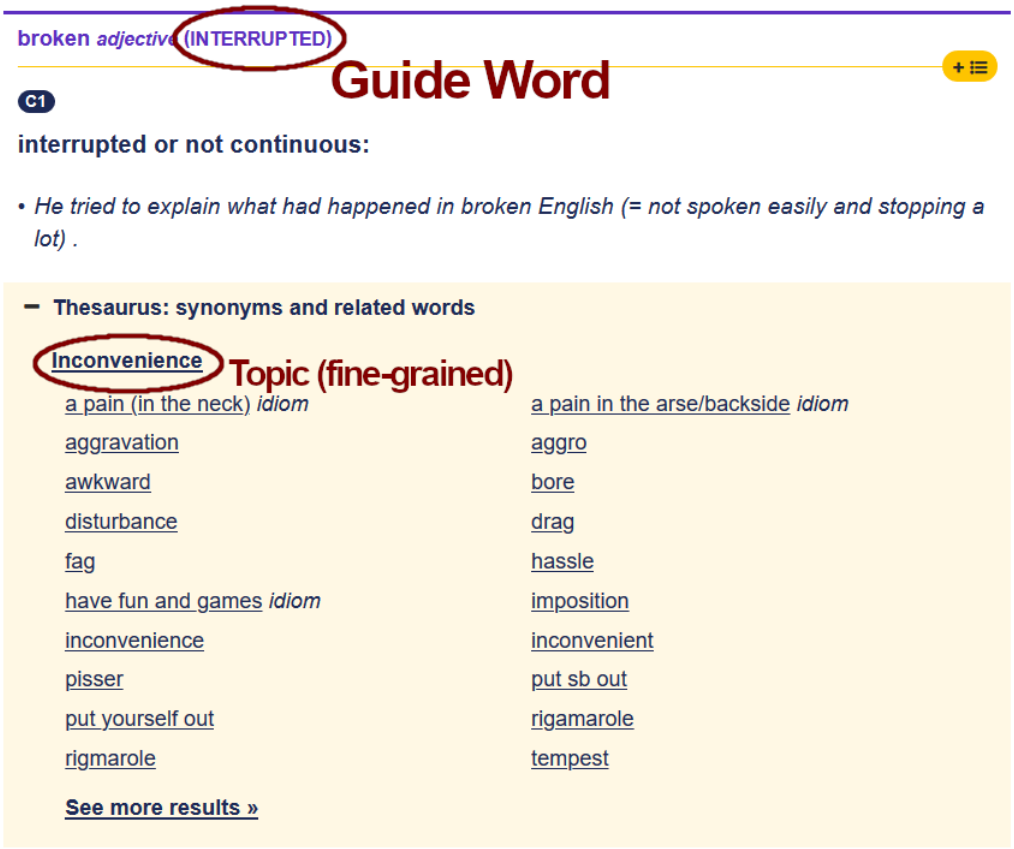

Figure 1: Screenshot of a CALD entry for *broken*. `Guide word` and `topic` are marked.

publicly accessible website provides definitions for these senses.[5]

To achieve the desired granularity of senses `guide words` and `topics` from the dictionary entries are used. According to the dictionary designers, `guide words` are supposed to help disambiguating entries when a word has more than one meaning. This means that the `guide words` are intended as simple and intuitive identifiers for meaning, usually expressed in one word. For example, the word *broken* has, amongst others, the `guide words` *damaged* and *interrupted*. Nothing equivalent to that is available for WordNet at this level of granularity. In addition, `topics` serve to provide further clustering for the words under the `guide word`: for example, the `topic` *inconvenience* clusters word senses associated with the `guide word` *interrupted* (see screenshot in Figure 1).

The `guide words` are more coarse-grained than the `topics`. For example, the word *head* is, amongst others, associated with the `guide word` *top part*, which in its turn subsumes no less than seven `topics`. These more fine-grained `topics` include *Edges and extremities of objects*, *Tools*, *Flowers - general words*, *Beer & cider*, *Parts of watercourses*, *Skin complaints & blemishes*, *Ahead, in front and beyond* (see Figure 2).

As a consequence of this coarseness difference, the sense annotation scheme is hierarchical with `guide words` as the higher level of annotation (see Table 3).

Guide Word	Topics
Know	• Sympathy & compassion • Empathy and sensitivity
Realize	• Linguistics: question words & expressions

Table 3: Available senses for *understand* in CALD

On the basis of the complexity data from Yimam et al. (2017), tokens with a sufficient number of different senses in the Cambridge Advanced Learner's Dictionary were selected. All tokens have at least three candidate senses at the coarser-grained `guide word` level. In addition, we eliminated all tokens with repeated `topics` to make sure that one can always identify one definition entry for the token in the online CALD.[6] For example, the sense of *close* with the `guide word` *not open* has two more fine-grained senses, one for closing a window, and the other for closing a shop, that are both associated with the `topic` *Closing and blocking*. Thus, the dataset poses an adequate challenge for word

[5]Cambridge University Press also provides an API service. For information on licensing and API see `https://dictionary-api.cambridge.org/`.

[6]The background CALD also uses internal entry identifiers, which are frequently the token, but sometimes distinguished by a number e.g. *head_1* for *head*. We note the internal entry identifier in square brackets as additional information, when it was available (see Figure 3).

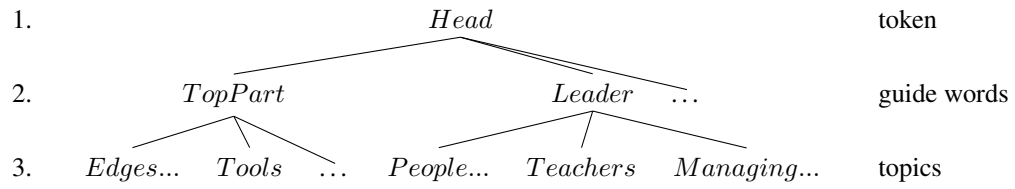

Figure 2: Part of sense hierarchy for the word *head*. *Top Part* and *Leader* are `guide words`, the `topics` are on the lower-level.

sense disambiguation software that can be evaluated using CALD.

The sense annotation was undertaken by four authors of this paper, all trained in linguistics and NLP. They had access to the tokens to be disambiguated, the full sentence contexts, and the online version of the CALD. The annotators were asked to select candidate senses from a drop-down menu, where the candidate senses had been produced automatically using the tokens to be disambiguated. The electronic CALD was searched for entries both for the word-form version of the word token and its lemmatized form. All entries found this way were available to the annotators for selection.

The annotators were advised to select a `guide word` and `topic` pair of the form "*GUIDE WORD* | *topic* [internal identifier]" whenever possible. The annotators, however, also had the option of choosing only a coarse-grained `guide word` if none of the available `topics` seemed appropriate. Thus, they could exploit the hierarchical nature of the annotation scheme to achieve better coverage (see Figure 3).

A special problem is created by multiword expressions (MWE). MWEs are combinations of two or more co-occurring words (e.g. *dry run*, *give up*, and *take into account*) which form idiosyncratic lexical units (Sag et al., 2002). WSD systems have difficulty assigning labels to individual words that are part of MWEs (Constant et al., 2017). In this work, we also take these cases into consideration. Specifically, following annotation guidelines, annotators first tried to find a relevant sense for a word token in CALD. If no appropriate sense could be found due to the word being part of an MWE, the word was annotated with the special label "PART OF MWE".[7]

For example, in the case of the MWE *common law* no sense of *common* available in the CALD is appropriate. This entry is annotated as "PART OF MWE". However, the word *care* in the context of the expression *take care of* is a case where the sense with the `guide word` *protection* can be simply selected. For any of these instances the MWE formation is noted in the comment slot, by first writing "MWE", then a pipe ("|"), and finally the whole MWE within which the token occurs (e.g. *MWE* | *common law* or *MWE* | *take care of*). Tokens which are part of Named Entities such as *bridge* in *Golden Bridge Terrace* are considered as a sub-type of MWEs and annotated the same way and with the same notation in this work.

The importance of incorporating MWE information in WSD has been explored in previous studies (Arranz et al., 2005; Finlayson and Kulkarni, 2011). Finlayson and Kulkarni (2011) reported a boost in performance of a WSD system when using an MWE detection strategy. Not only does MWE information help WSD, but also, we believe that their idiosyncratic behaviour calls for special treatment in order to thoroughly investigate sense complexity.

4. Statistics

Type	Count
Tokens	1432
Types of token	211
Average number of senses	4.7
Average context length	31.7

Table 4: Statistics describing the dataset. "Tokens" stands here for tokens to be disambiguated.

As mentioned above, the tokens have been selected for expressing at least three possible senses at the coarse-grained level of `guide words`. This led to 1423 tokens with candidate senses, for which there are overall 6699 candidate senses available at the fine-grained level of `topics`, which amounts to the average of 4.7 senses per token. The maximum number of senses is 18, for the word *head*.

By contrast, WordNet provides 13236 synsets and does not provide a more coarse-grained level equivalent to the `guide word` level. The maximum number of synsets per word is 72, for the token *broken*.[8] By comparison, there are only 5 senses for *broken* in the CALD version. Thus, it can be seen that our approach provides more coarse-grained word sense as intended.

In total, there are 211 types of tokens.[9] The average length of the context, including the token to be disambiguated and punctuation marks, is 31.7 tokens. The context is typically, but not always, one sentence.

4.1. Inter-Annotator Agreement

To determine annotator reliability we calculate the Fleiss' Kappa κ (Fleiss, 1971) for groups of words with the same number of sense categories, and then take a weighted average of these values:

[7] Not all components of an MWE necessarily deviate from their established senses. We acknowledge the appropriate senses listed in CALD and annotators choose the right sense if available. At the same time, we have a comment section for annotators to write down the MWE which the word belongs to.

[8] As in the case of the CALD senses, this count includes the senses found by looking for the unchanged token and the lemmatized form.

[9] This count includes different grammatical forms of the same word, e.g. *countries* and *country*.

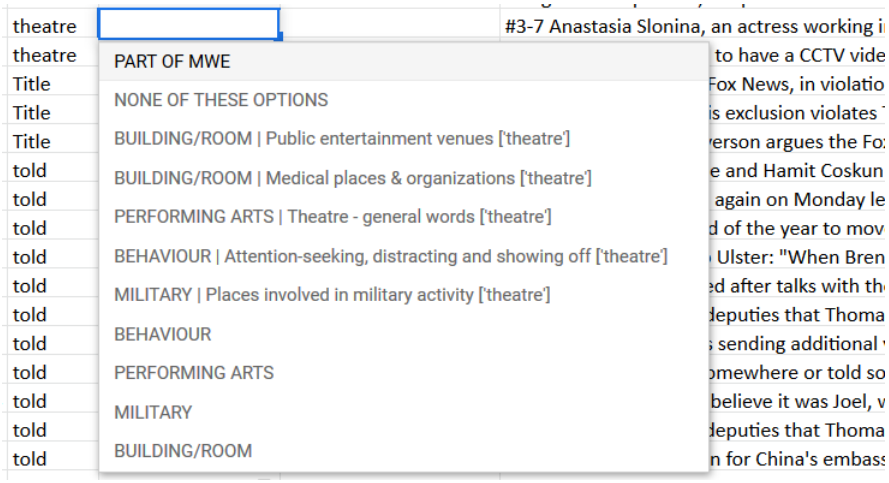

Figure 3: Screenshot of annotation options presented to word sense annotators for the token *theatre*. Capitalised words are `guide words`, followed by `topic` after the pipe. Words in square brackets represent entry name in CALD.

$$\overline{\kappa} = \frac{\sum_{c \epsilon C} |c| \kappa_c}{\sum_{c \epsilon C} |c|} \quad (1)$$

where κ_c represents the Fleiss' kappa agreement of annotators for each group, containing words with the same number of sense categories c, and $|c|$ represents the number of examples within the group. Over one-third of the dataset was labelled by all four annotators (501 instances in total), with the remaining 992 phrases annotated by three annotators each. We calculate the agreement for these groups separately using the aforementioned technique. The average value obtained for 4-way agreement was $\overline{\kappa} = 0.681$, and $\overline{\kappa} = 0.646$ in cases annotated by 3 annotators. Both values constitute substantial agreement according to Landis and Koch (1977). We additionally calculate the pairwise percentage agreement across annotators, which equals to 81.93%, as well as the overall percentage of complete agreement, which equals to 69.4%. We note that in most cases the disagreement is due to the existence of subtle sense distinctions: for instance, the general category *MONEY* vs. the category *MONEY | Profits & losses*.

5. Conclusion

SeCoDa is a dataset that combines sense and complexity information. We make the dataset publicly available to enable research into the complexity differences at the level of senses rather than tokens.

We adopted a hierarchical sense annotation scheme drawing on information available in the Cambridge Advanced Learner's Dictionary. This scheme provides more coarse-grained senses than WordNet.

Acknowledgements

We thank Ted Briscoe for advice and support. We also thank the three anonymous reviewers of this paper. This paper reports on research supported by Cambridge Assessment, University of Cambridge.

6. Bibliographical References

Agirre, E. and Lacalle, O. L. D. (2003). Clustering Wordnet Word Senses. In *Proceedings of the Conference on Recent Advances on Natural Language (RANLP'03)*.

Arranz, V., Atserias, J., and Castillo, M. (2005). Multiwords and word sense disambiguation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 250–262. Springer.

Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Council of Europe. (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

Edmonds, P. and Cotton, S. (2001). SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France, July. Association for Computational Linguistics.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, Mass.

Finlayson, M. A. and Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gooding, S. and Kochmar, E. (2018). Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Gooding, S. and Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Ide, N. and Wilks, Y. (2007). Making Sense about Sense. In Eneko Agirre et al., editors, *Word Sense Disambiguation: Algorithms and Applications*, pages 47–73. Springer, New York, November.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Levinson, S. C. (1995). Three Levels of Meaning. In F. R. Palmer, editor, *Grammar and Meaning*, pages 90–115. Cambridge University Press, Cambridge.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Language, Speech, and Communication. MIT Press, Cambridge, Mass.

Litkowski, K. C. (2004). SENSEVAL-3 Task: Word-Sense Disambiguation of WordNet Glosses. In *In Proc. of SENSEVAL-3 Workshop on Sense Evaluation, in the 42th Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.

McCarthy, D. (2006). Relating Wordnet Senses for Word Sense Disambiguation. In *Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, April.

Mihalcea, R. and Moldovan, D. I. (2001). EZ.WordNet: Principles for Automatic Generation of a Coarse Grained WordNet. In *FLAIRS Conference*.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69, February.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making Fine-Grained and Coarse-Grained Sense Distinctions, Both Manually and Automatically. *Natural Language Engineering*, 13(2):137–163, June.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, W., Peters, I., and Vossen, P. (1998). Automatic Sense Clustering in Eurowordnet. In *Proceedings of LREC'1998*.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). CWIG3G2 - complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yuan, D., Richardson, J., Doherty, R., Evans, C., and Altendorf, E. (2016). Semi-supervised Word Sense Disambiguation with Neural Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385, Osaka, Japan, December. The COLING 2016 Organizing Committee.