# Postersession der Computerlinguistik/Postersession of the Section on Computational Linguistics

## Trafilatura: An Open-Source Tool for Web Corpus Construction

Adrien **Barbaresi**, Berlin-Brandenburgische Akademie der Wissenschaften
barbaresi@bbaw.de

Large “offline” web text collections are now standard among the research community in linguistics and natural language processing. The construction of such corpora notably involves “crawling, downloading, ‘cleaning’ and de-duplicating the data, then linguistically annotating it and loading it into a corpus query tool” (Kilgarriff 2007). Although text is ubiquitous on the Web, extracting information from web pages can prove to be difficult. Web documents come in different shapes and sizes mostly because of the wide variety of genres, platforms, and content management systems, and not least because of greatly diverse publication goals. Web crawling involves a significant number of design decisions and turning points in data processing, without which data and applications turn into a “Wild West” (Jo & Gebru 2020). Researchers face a lack of information regarding the content, whose adequacy, focus, and quality are the object of a post hoc evaluation (Baroni et al. 2009). Comparably, web corpora usually lack metadata gathered with or obtained from documents. Between opportunistic and restrained data collection, a significant challenge lies in the ability to extract and pre-process data to meet scientific expectations with respect to corpus quality.

*Trafilatura* is a library and command-line tool used for corpus construction within the lexicographic information platform *dwds.de* (Geyken et al. 2017) which hosts and provides access to a series of metadata-enhanced web corpora (Barbaresi 2016). It seamlessly downloads, parses, and scrapes web page data. It handles the extraction of metadata, main body text and comments while preserving parts of the text formatting and page structure. Link discovery in feeds and sitemaps is also included. The output is then converted to common formats (TXT, CSV, JSON, XML & XML-TEI). Distinguishing between a whole page and the page's essential parts helps to alleviate many quality problems by dealing with the noise caused by recurring elements (headers and footers, ads, links/blogroll, etc.), so that the software both facilitates text data collection and enhances corpus quality. As evaluations of extraction tools show significant domain-related disparities (Barbaresi & Lejeune 2020), the experiments at hand show that the tool performs better than known alternatives. It is freely available under an open-source license: https://github.com/adbar/trafilatura

### References

Barbaresi, Adrien, 2016. "Efficient construction of metadata-enhanced web corpora". In Paul Cook, Stefan Evert, Roland Schäfer, and Egon Stemle, eds., *Proceedings of the 10th Web as Corpus Workshop*, Association for Computational Linguistics*,* 7-16.

Barbaresi, Adrien, and Gaël Lejeune, 2020. "Out-of-the-Box and into the Ditch? Multilingual Evaluation of Generic Text Extraction Tools", *Proceedings of the 12th Web as Corpus Workshop (LREC 2020)*, ELRA, 5-13.

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano, and Eros Zanchetta, 2009. "The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation*, 43(3): 209-226.

Geyken, Alexander et al., 2017. “Die Korpusplattform des Digitalen Wörterbuchs der deutschen Sprache (DWDS)”, *Zeitschrift für germanistische Linguistik*, 45(2): 327-344.

Jo, Eun Seo, and Timnit Gebru, 2020. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning", In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 306-316.

Kilgarriff, Adam, 2007. "Googleology is bad science", *Computational Linguistics*, 33(1): 147-151.

## Advancing Neural Question Generation for Formal Pragmatics

Speaker 1 (Kordula **De Kuthy**, Universität Tübingen), speaker 2 (Haemanth **Santhi Ponnusamy,** Universität Tübingen), speaker 3 (Madeeswaran **Kannan**, Universität Tübingen), speaker 4 (Detmar **Meurers**, Universität Tübingen), {kdk,mkannan,hsp,dm}@sfs.uni-tuebingen.de

Question generation, creating questions for a given sentence or paragraph, is a challenging task with many applications, from question answering, via dialogue systems, to reading comprehension tasks. The recent state-of-the-art approaches are generally based on neural networks. The task of QG is typically formulated as a sequence-to-sequence (seq2seq) learning problem in which a sentence is mapped to a corresponding question (cf., e.g., Pan et al., 2019).

In formal pragmatics, questions also play an prominent role in so-called Questions-under-Discussion (QuD, Roberts, 2012) approaches. Questions there make explicit the interface between the information structure of a sentence and the discourse structure that the sentence functions in. Under such a QuD perspective, for every sentence in a text, a question needs to be formulated – and indeed explicit guidelines have been defined to support reliable manual QuD annotation (Riester et al., 2018). De Kuthy et al. (2020) argue that such question generation should be automated for the analysis of large corpora, and they propose a seq2seq neural network approach to generate all potential questions for a given sentence. They show that the approach learned to (often) predict the correct question word for a given answer and generated questions that correctly reflect the word order properties of questions in German.

There are, however, clear challenges for such a seq2seq architecture that generates questions for any type of data set. One problem are rare or unknown words that have to be predicted. In most neural generation architectures, words are the basic tokens. Pretrained word embeddings are used to initialize the token embedding matrix with a fixed vocabulary. In any corpus material serving as input there are likely to be rare or unknown words that are not part of the fixed vocabulary and therefore cannot be predicted in the output, the generated question. This indeed is a major issue in De Kuthy et al.'s question generation approach. To overcome this problem, they implemented an ad-hoc post-processing step: Each generated question is checked for markers indicating the places where an OOV token appears. A heuristic then tries to identify that missing word in the source sentence and insert it in the output.

Here we propose to adopt a pointer-based neural architecture for QG. We show that such an architecture is more successful than the seq2seq based model, replacing the post-processing step used in De Kuthy et al. (2020) into a design feature of the neural architecture. Architecturally separating the copying from the generation component also readily supports the integration of linguistic information needed to determine the question phrase to be generated. Furthermore, the pointer-based architecture is able to generalize the task of question generation in identifying the material that is identical between source sentence and question and that can simply be copied over. A quantitative evaluation using BLEU scores and an in-depth qualitative evaluation show that indeed the pointer-based model with additional linguistic features is the best system for the task of generating questions to advance discourse analysis.

### References

De Kuthy, Kordula, Madeeswaran Kannan, Haemanth Santhi Ponnusamy & Detmar Meurers. 2020. *Towards automatically generating questions under discussion to link information and discourse structure.* In Proceedings of the 28th international conference on computational linguistics, Barcelona, Spain.

Pan, Liangming, Wenqiang Lei, Tat-Seng Chua & Min-Yen Kan. 2019. *Recent advances in neural question generation.* arXiv preprint arXiv:1905.08949 .

Riester, Arndt, Lisa Brunetti & Kordula De Kuthy. 2018. *Annotation guidelines for questions under discussion and information structure.* In E. Adamou, K. Haude & M. Vanhove (eds.), Information structure in lesser-described languages: Studies in prosody and syntax Studies in Language Companion Series, John Benjamins.

Roberts, Craige. 2012. *Information structure in discourse: Towards an integrated formal theory of pragmatics.* Semantics and Pragmatics 5(6). 1–69. doi: 10.3765/sp.5.6.

## Recognizing deliberate metaphors

Stefanie **Dipper**, Anna **Ehlert**, Doreen **Scholz**, Felix **Theodor**, Larissa **Weber** (Ruhr-Universität Bochum)
stefanie.dipper@rub.de, anna.ehlert@rub.de, doreen.scholz@rub.de, felix.theodor@rub.de, larissa.weber@rub.de

Metaphors are a widespread phenomenon that occurs frequently in various types of text. Metaphors involve a "mapping across two conceptual domains" (Steen, 2007): they refer to the properties of one concept in order to describe and clarify the properties of another concept. For example, in (1) the concept "skeleton" is used to refer to the function of a skeleton as a supporting structure of a body, which can be transferred to the supporting structure of buildings. In addition, the concept "skeleton" indicates that the associated body is no longer alive, otherwise the skeleton would not be visible at all and would not be able to rise up into the air. transferred to buildings this means that the buildings are destroyed.

(1) Skeletons of skyscrapers rose into the sky.

Many modern linguistic expressions go back to metaphors, which have now been conventionalized, however. In (2), for example, the concept "attack" is no longer associated with a warlike activity, but is directly understood as a form of argumentation.

(2) Lakoff attacked Glucksberg.

Especially metaphors of type (1), which we call "deliberate metaphors", pose a challenge for automatic processing, because certain expressions are not used literally or with a non-canonical meaning.

In our poster we want to present our work on the automatic recognition of deliberate metaphors. We present our annotation guidelines as well as results of a corpus of sermons currently being annotated according to these guidelines. Furthermore, we implement a recognizer based on the approach of Shutova et al. (2012), but adapting and extending it to German.

### References

Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical Metaphor Processing, *Computational Linguistics*, 39(2): 301–353.

Steen, Gerard J. 2007. Finding Metaphor in Discourse: Pragglejaz and Beyond. *Cultura, Lenguaje y Representación*, 5: 9–25.

## Annotating and interpreting deliberate metaphors: An implementation of Steen’s Five Step Method

Stefanie **Dipper**, Frederik **Elwert**, Tim **Karis** (Ruhr-Universität Bochum)
stefanie.dipper@rub.de, frederik.elwert@rub.de, tim.karis@rub.de

The main characteristic feature of metaphors is the mapping from one conceptual domain to another, with the goal of appropriately describing the concept of the target domain using the concept of the source domain. Metaphor fulfils a special role in religious language, where its capacity to express ideas about an abstract entity with reference to a well-known concrete entity works as a means to make statements about the transcendent. In (1), an extract from a religious text in Middle High German, the metaphor SALVATION IS HEALING is used to convey religious ideas: abstract theological notions such as original sin and salvation are mapped onto a more tangible domain by referring to the concepts of wounding and healing.

(1) so vnsir herre got alle die wnden virbindit die wir íe von adames svndon gefrvmeton
‘Thus our Lord God binds up all the wounds we have suffered through Adam’s sin.’

We distinguish between two steps in metaphor analysis: metaphor identification and metaphor interpretation. For the first steps, there are comprehensive guidelines (MIP, Pragglejaz Group 2007, and MIPV, Steen et al., 2010). For the second step, Steen (2007) proposed the ‘Five Step Method’. In our poster, we present an implementation and extension of Steen's method that supports annotators in identifying and writing up explicitly stated propositions as well as implicit assumptions that are relevant and necessary to arrive at the metaphor's interpretation.

### References

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1): 1–39.

Steen, Gerard J. 2007. Finding Metaphor in Discourse: Pragglejaz and Beyond. *Cultura, Lenguaje y Representación*, 5: 9–25.

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of Converging Evidence in Language and Communication Research. John Benjamins Publishing Company, Amsterdam/Philadelphia.

# Machine learning approaches to analyzing German synthetic compounds

Carlotta J. **Hübener**, Universität Hamburg
carlotta.huebener@uni-hamburg.de

Synthetic compounding (e.g., *schönheitsliebend* ‘beauty-loving’) is a highly productive word-formation pattern in German (Neef 2015: 588), which can give insight into the interplay of word-formation and syntax. However, little is known about the internal argument structure of the words it yields. The poster presents an application of a machine learning model for analyzing the syntactic structure of synthetic compounds, focusing on noun-participle combinations.

There is a consensus in the literature that most synthetic compounds are based on accusative phrases (e.g., *ekelerregend* ‘nauseating’ ~ *Ekel*$_{\mathrm{ACC}}$ *erregen* ‘to arouse disgust’). It is unclear to what extent there is a correspondence to dative and genitive phrases as well (e.g., *zweckentsprechend* ‘appropriate’ ~ *Zweck*$_{\mathrm{DAT}}$ *entsprechen* ‘to correspond to the purpose’). Knowing this distribution is important, however, for instance when investigating word-formation restrictions and the interplay of grammar and the lexicon in general.

To examine the internal argument structure of synthetic compounds, the valencies of their base verbs have to be determined. For this purpose, automatic dependency parsing is advantageous: Large numbers of texts can be parsed within a short time and at low cost. The present study used a pretrained dependency parsing model from the Python library spaCy (Honnibal & Montani 2017) to identify and classify clause constituents. In a first step, a list of noun-participle combinations was extracted from the DWDS core corpus and the DIE ZEIT corpus. The corresponding noun-verb combinations were queried in the DWDS core corpus (e.g., *Kopf* ‘head’ and *schütteln* ‘to shake’ for *kopfschüttelnd* ‘head-shaking’). Then, the dependency parser analyzed the syntactic dependency between noun and verb. For instance, the model identified 1,705 sentences with a syntactic dependency between the lexemes *Kopf* and *schütteln*, classifying this relation as “oa” (accusative object) in 98.3% of cases. Thus, *kopfschüttelnd* is obviously based on an accusative phrase.

With manually annotated data serving as a reference standard, the approach achieved a micro-average accuracy of 0.94 (average precision: 0.99, average recall: 0.89, average $F_1$ score: 0.94) for a sample of 404 noun-participle combinations. Restricted to well-attested verbal phrases in the corpus (*f* > 10), the accuracy increased to 0.97. Both the manually and the automatically annotated data confirm that most noun-participle combinations correspond to accusative phrases (99.5% or 94%, respectively). The results suggest that spaCy’s dependency parser is an overall reliable tool offering promising possibilities for the further examination of synthetic compounds, for instance regarding the relationship between grammar and the lexicon.

## References

Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. DIE ZEIT corpus. https://www.dwds.de/d/korpora/zeit.

Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. Core corpus. https://www.dwds.de/d/k-referenz#kern.

Honnibal, Matthew, and Montani, Ines. 2017. *spaCy 2. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

Neef, Martin. 2015. "33. Synthetic compounds in German." In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, eds. *Word-formation. An international handbook of the languages of Europe.* Berlin: de Gruyter. 581-592.

# „Ich gehe kurz Zigaretten holen“- Diskurs berechnen mit Word Embedding.

Speaker 1 (Zakharia **Pourtskhvanidze**, Institut für Empirische Sprachwissenschaft, Goethe-Universität Frankfurt) pourtskhvanidze@em.uni-frankfurt.de

0. **Grundannahme**: (1) Die Diskurse lassen sich mithilfe von korpuslinguistischen Tools grundsätzlich berechnen. (Bubenhofer 2008); (2) Die Satzhypostasen sind Sprachgebrauchsmuster und indizieren die spezifischen korrespondierenden Diskurse.
1. **Ontologie des Zigarettenholen-Diskurses im deutschsprachigen Gebrauch**.

*Heute will ich Ihnen erzählen, wie das damals wirklich war, als ich „mal eben Zigaretten holen ging“ und erst 23 Jahre später an meinem Heimatort zurückkehrte…* (Schottleitner).

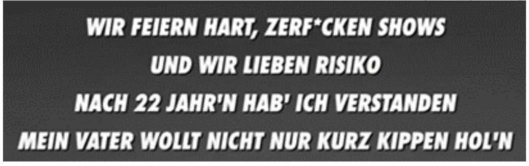
*"Ich geh mal Zigaretten holen"-Fälle. Hallo ihr! Wisst ihr, ob es wirklich so viele Männer gibt bzw. gegeben hat, die gesagt haben "ich gehe mal kurz Zigaretten holen" oder so etwas Ähnliches und dann einfach verschwunden sind?* (Brigitte).

BIN KURZ ZIGARETTEN HOLEN.

KOMME IN ZIRKA ZEHN JAHREN ZURÜCK.

"ich gehe mal eben zigaretten holen..."

"und wozu brauchst du die reisetasche?"



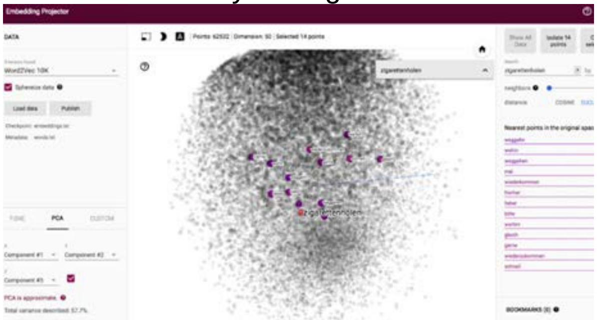
WIR FEIERN HART, ZERF*CKEN SHOWS
UND WIR LIEBEN RISIKO
NACH 22 JAHR'N HAB' ICH VERSTANDEN
MEIN VATER WOLLT NICHT NUR KURZ KIPPEN HOL'N

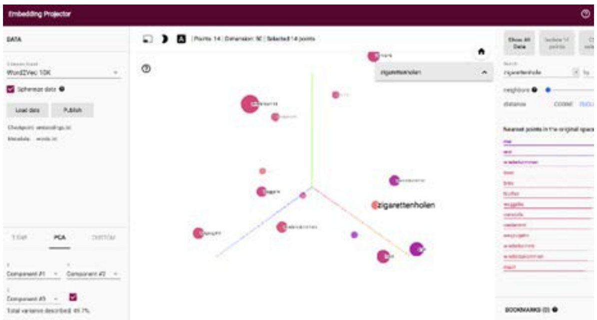
2. **Problemstellung.** (1) Überprüfung der Schlagwortfähigkeit der Satzhypostase „Ich-gehe-nur-mal-kurz-Zigaretten-holen“, (2) Ermittlung von korrespondierenden Diskurse im korpuslinguistischen verfahren (Word Embedding).
3. **„Zigarettenholen“ Zwischen Word- und Phrase Embedding**.

Es wurde ein Webkorpus aus ca. 1Mio. Token verwendet. Im 50-dimensionalen Tensor-Raum wurden die Vektorendaten analysiert und 14 miteinander zusammenhängende Knoten isoliert. Die verschiedenen phrastischen Versionen des Satzes wurden in einem Uni-Gram „Zigarettenholen“ umgewandelt und der Vektor dieses kumulativen Uni-Grams wurde in der Analyse eingesetzt.

Durch das Netzwerk-Modell, dass aufgrund der Word Vectors Analyse entsteht, lässt sich eine allgemeine Diskurs-Klassifizierung vollziehen, in dem die folgenden Diskursschlag-wörter errechnet wurden:

1. 'warten'
2. 'geblufft'
3. 'weggehn'





4. **Ergebnis.** Die Konstruktion „Ich-gehe-nur-kurz-Zigaretten-holen“ hat sich mit der konzeptuellen Bedeutung „weggehen ohne sich zu verabschieden und/oder Gründe zu nennen“ im Beziehungs- bzw. Liebes-Diskurs verselbstständigt. Die Konzepte wie {„weggehn“, „bluff“, „wieder(zu)kommen“} sind im Word-Embedding-Verfahren errechnet und stehen als korrespondierenden Diskurs-Schlagwörter.

## References

Bubenhofer, Noah. 2008. Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse. In: Spitzmüller/Warnke. Methoden der Diskurslinguistik: sprachwissenschaftliche Zugänge zur transtextuellen Ebene. Berlin / New York: de Gruyter, 407-434.

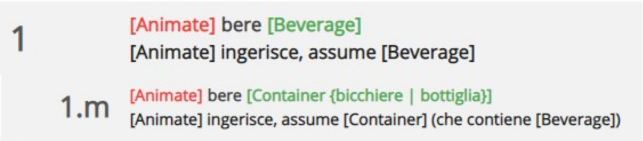
Schottleitner, Vera. „Männer u.s.w.“ Gedichte und Essays. http://www.literaturlinie.de/Nicht_wahr.pdf. Eingesehen am 01.04.2020.

# Annotating Metonymic Relations in a Corpus-based Resource for Italian Verbs

Emma **Romani**, University of Pavia (Italy), Elisabetta **Ježek**, University of Pavia (Italy)
emma.romani01@universitadipavia.it, jezek@unipv.it

In this poster we address the results of a research thesis (Romani 2020) dedicated to the annotation of metonymies. Metonymy is the language phenomenon for which one referent is used to denote another referent associated with it. In our research, we investigated metonymy from a corpus-based perspective, through the analysis of corpus data and an annotation performed in T-PAS, a corpus-based resource for Italian verbs (Ježek et al. 2014). T-PAS consists in a repository of Typed Predicate Argument Structures (called *t-pas* or pattern, one for each meaning of each verb), i.e., verbal patterns with arguments signalled by semantic types, linked to manually annotated corpus instances.

The annotation of metonymies was performed starting from a list of 30 verbs contained in T-PAS. Our work was intended as an implementation of the resource; specifically, we annotated corpus instances of the verbs containing metonymies and created metonymic sub-patterns linked to them (Fig. 1). We followed a corpus-based methodology, which was also devised to distinguish metonymies from complex types (Ježek & Vieu 2014).



1 [Animate] bere [Beverage]
[Animate] ingerisce, assume [Beverage]

1.m [Animate] bere [Container {bicchiere | bottiglia}]
[Animate] ingerisce, assume [Container] (che contiene [Beverage])

Fig. 1. Metonymic sub-pattern for t-pas 1 of the verb *bere* (‘to drink’) in T-PAS

We also conceived a theoretical framework to represent the metonymies found through the corpus analysis, by designing a map and by compiling a list of the metonymic *relations* occurring in the verbal patterns (in case of acceptance, the map and the list will be included in the poster). A *relation* is a brief description that illustrates how the *metonymic semantic type* is connected to the *target semantic type*; for example, [Container] (metonymic semantic type) ‘contains’ (the *relation*) [Beverage] (target semantic type). Both the map and the list depict the complexity and variety of the phenomenon, in terms of number of possible metonymic relations and of the semantic types interested.

In future perspectives, we intend to enrich the map and the list with new relations by extending the number of verbs investigated and to evaluate the annotation procedure. We are also interested in a crosslinguistic comparison of our results with those in the Croatian sister project of T-PAS (CROATPAS, Marini & Ježek 2020). The annotated corpus data, as well as the relations, will be useful for automatic detection of metonymies (Markert & Nissim 2009). To our knowledge, little work has been done on this for Italian language: it will be therefore intriguing to test our data in NLP tasks.

## References

Ježek, Elisabetta, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. “T-PAS: A resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing”. *Proceedings of LREC*. 890-895.

Ježek, Elisabetta, and Laure Vieu. 2014. “Distributional Analysis of Copredication: Towards Distinguishing Systematic Polysemy from Coercion”. *Proceedings of the First Italian Conference on Computational Linguistics*. 219-223.

Marini, Costanza, and Elisabetta Ježek. 2020. “Annotating Croatian Semantic Type Coercions in CROATPAS”. *Proceedings of the 16th Joint ACL-ISO Workshop ISA-16*, pp. 49–58.

Markert, Katia, and Malvina Nissim. 2009. “Data and models for metonymy resolution”. *Lang Resources & Evaluation*. 43: 123–138.

Romani, Emma. 2020. *Searching for Metonymies in Natural Language Texts. A Corpus-based Study on a Resource for Italian Verbs*. BA Thesis, Pavia: University of Pavia.

# Easy and Reproducible WebAnno Project Management

Adam **Roussel**, Ruhr-Universität Bochum
roussel@linguistics.rub.de

In large annotation projects and in educational settings, you may need to create a large number of WebAnno (Eckart de Castilho et al., 2016) projects and/or user accounts at once, which can be tedious and time-consuming to do by hand. And, especially where it is important that the projects be configured in a particular way, it can also be error-prone to do everything with the graphical interface.

With this poster I'd like to introduce PyWebAnno, a Python script that helps you orchestrate collections of WebAnno projects and users. You can generate a large number of user accounts, notify these users of their login data by email, assign them automatically to WebAnno projects, and then, when the course has concluded, remove the generated projects and users – but only the right projects and users, leaving other projects on your WebAnno instance untouched. With PyWebAnno, you can also specify the documents that should belong in each project or have these assigned automatically. Finally, with the facility of uploading annotations to the generated projects, you can use the Curation function to compare the students' annotations with a gold-standard or automatically-generated annotations.

PyWebAnno is free and open source software available at:

https://git.noc.rub.de/ajroussel/pywebanno

## References

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016): A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan

## Differences between German and English Text Simplification

Regina **Stodden**, Heinrich-Heine-University Düsseldorf
regina.stodden@uni-duesseldorf.de

Text simplification is a natural language processing task which aims at automatically reducing the complexity of a given text. This research area is part of natural language generationand (monolingual) machine translation. Text simplification focus on generating a more easily understandable version of a complex text for language learners or low literacy. The simplified text should preserve the meaning of the complex text and should not contain grammatical errors (Alva-Manchego et al., 2020). So far, text simplification research mostly focuses on English (see Alva-Manchego et al. (2020) for an extensive list), and only a few studies exist for German (Klaper et al., 2013; Battisti et al., 2020; Mallinson et al., 2020).

The German text simplification research can nowadays benefit from an active community in easy-to-read German, including translation offices related to practices and research facilities related to theory. Two main versions exist of German easy-to-read languages, i.e., plain language (de: “Einfache Sprache”) and easy language (de: “Leichte Sprache”) (Maaß, 2020). Plain language seems more applicable to text simplification than easy language because the overall variant and its complexity are closer to everyday German. In a content analysis of recommendations on how to write German plain language and text simplification research papers, we found items that are more relevant in English than German and vice versa. These items specify the transformations during a simplification, e.g., substituting complex words or deleting superfluous information.

Both areas agree on deleting or replacing complex words and sentence splitting. In comparison to easy-to-read English, German plain language focuses more on compound splitting and compound segmentation. Furthermore, German plain language recommendations contain more frequent changes in the verb’s voice, deletions of phrases and clauses, and explanations of complex words in a new sentence. In contrast, in text simplification research, sentence reordering is mentioned more often than in German plain language.

On the poster, we will explain more briefly text simplification and the differences between German plain language and German easy language. Furthermore, we will present text simplification transformations that are specific for German and English and give examples for the transformations in both languages.

### References

Alva-Manchego, Fernando, Carolina Scarton, and Lucia Specia. 2020. “Data-driven sentence simplification: Survey and benchmark.” *Computational Linguistics, 46(1)*:135–187.

Battisti, Alessia, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. “A corpus for automatic readability assessment and text simplification of German.” In *Proceedings of the 12th LREC*, pages 3302–3311, Marseille, France. ELRA.

Klaper, David, Sarah Ebling, and Martin Volk. 2013. “Building a German/simple German parallel corpus for automatic text simplification.” In *Proceedings of the 2ndWorkshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria. ACL.

Maaß, Christiane. 2020. “Easy Language - Plain Language - Easy Language Plus. Balancing Comprehensibility and Acceptability.” *Easy – Plain – Accessible*. Frank & Timme, Berlin.

Mallinson, Jonathan, Rico Sennrich, and Mirella Lapata. 2020. “Zero-shot cross-lingual sentence simplification.” In *Proceedings of EMNLP 2020*, pages 5109–5126, Online. ACL.

# Cross-Lingual Word Embeddings for Extremely Low-Resource Languages: Improving Bilingual Lexicon Induction for Hiligaynon

Speaker 1 (Mary Ann **Tan**), Co-Author 1 (Dario **Stojanovski**, CIS LMU), Co-Author 2 (Alexander **Fraser**, CIS LMU)
anntanp@gmail.com, dario@cis.lmu.de, fraser@cis.lmu.de

Cross-Lingual Word Embeddings (CLWEs) have been experiencing a surge in popularity in the past couple of years due to the remarkable progress in machine learning techniques, the availability of large natural language processing (NLP) datasets and the exponential growth in computing power. CLWEs represent words from several languages in a shared embedding space; a more standard bilingual representation is called Bilingual Word Embeddings (BWEs). This research area has gained traction in the field of machine translation (MT) primarily because of its application to the task of Bilingual Lexicon Induction (BLI), which uses BWEs to learn word-pair translations with no or little supervision.

However, as with most research areas in NLP, progress is mostly limited to resource-rich Indo-European languages. Recent work on English (EN) and Hiligaynon (HIL), an extremely low-resource language and the 4th most spoken native language in the Philippines (10 million speakers), did not manage to produce BWEs of reasonable quality primarily due to a lack of a sizable monolingual corpus (Michel et al., 2020).

Mapping-based approaches to CLWEs have prevailed due to their simplicity, computational tractability and relaxed data requirements (Mikolov et al., 2013; Faruqi and Dyer, 2014; Dinu et al., 2015; Lazaridou et al., 2015; Xing et al., 2015, Artetxe et al., 2016). This approach requires only two (2) monolingual word embeddings (MWEs), pre-trained separately on large unannotated monolingual corpora, and a seed lexicon containing word pairs from the source and the target language. Its objective is to project the word embeddings of the source MWEs to the embedding space of the target MWEs by learning a transformation matrix using the seed lexicon as its bilingual supervision.

Previous studies on low-resource languages achieved zero or close to zero precision-at-1 (P@1) with EN-HIL (Michel et al., 2020), and a collection of other non-heterogeneous BWEs trained on 5M token corpora (Dyer, 2019). In this study, we showed that EN-HIL BWEs, trained on a target corpus containing just a little over 1M tokens, yielded a BLI performance of P@1 at 9.26%. This was achieved by adapting an iterative orthogonal mapping with generative adversarial approach (Conneau et al., 2018), by properly curating the seed lexicon and by employing resource-rich languages as pivots for transfer learning. The pivot languages used for our experiments were two (2) Philippine languages, Filipino and Cebuano, another Austronesian language, namely Bahasa Indonesia, and Spanish, a major source of foreign loan words in Hiligaynon (Kaufmann, 1934). Among the pivot languages used, Spanish performed best due to the high quality of its MWEs.

…

## References

Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko. 2016. “Learning principled bilingual mappings of word embeddings while preserving monolingual invariance.” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Conneau, Alexis, Lample, Guillaume, Ranzato, Marc'Aurelio, Denoyer, Ludovic, and Jégou, Hervé. 2018. “Word translation without parallel data.” In Proceedings of the 6th International Conference on Learning (ICLR 2018).

Dinu, Georgiana, Lazaridou, Angeliki, and Baroni, Marco. 2015. “Improving zero-shot learning by mitigating the hubness problem.” In Proceedings of ICLR (Workshop Track).

Dyer, Andrew R. 2019. “Low supervision, low corpus size, low similarity! Challenges in cross-lingual alignment of word embeddings: An exploration of the limitations of cross-lingual word embedding alignment in truly low resource scenarios.” Master’s Thesis, Department of Linguistics and Philology, Uppsala University.

Faruqui, Manaal and Dyer, Chris. 2014. “Improving vector space word representations using multilingual correlation.” In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Kaufmann, John. 1934. “Visayan-English Dictionary.” CreateSpace Independent Publishing Platform.

Lazaridou, Angeliki, Dinu, Georgiana, and Baroni, Marco. 2015. “Hubness and pollution: Delving into cross-space mapping for zero-shot learning.” Association for Computational Linguistics, volume 1 (Long Papers), pages 270–280.

Michel, Leah, Hangya, Victor, and Fraser, Alexander. 2020. “Exploring bilingual word embeddings for hiligaynon, a low-resource language.” In Proceedings of the 12th Conference of Language Resources and Evaluation (LREC), pages 188-193. European Language Resources Association.

Mikolov, Tomas, Le, Quov. V., and Sutskever, Ilya. 2013. “Exploiting similarities among languages for machine translation.” https://arxiv.org/abs/1309.4168

Xing, Chao, Wang, Dong, Liu, Chao, and Lin, Yiye. 2015. “Normalized word embedding and orthogonal transform for bilingual word translation.” In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

# A New System for Rewriting Linguistic Annotations

Mark-Matthias **Zymla**, University of Konstanz
Mark-Matthias Zymla@uni-konstanz.de

We present a new system for expanding and rewriting linguistic annotations. As an example,we focus on the application of this system to syntactically annotated data. We showcase how the system can be used to add semantic annotations to some (syntactic) input and how it can be integrated into an annotation pipeline to produce semantic representations.

The present system was primarily inspired by the packed rewrite system (PRS) contained in the Xerox Linguistics Environment (XLE; Crouch et al. (2017)). The PRS has been successfully used to implement large-scale semantic parsing and even semantic reasoning (Bobrow et al., 2007). However, the system is not supported by XLE anymore and is not publicly available. We provide a new take on the PRS that aims to make the system available and useful to a broader audience. For this, inspiration is drawn from recent work in linguistic annotation. Ide and Bunt (2010) pointed out that linguistic annotations share a common core that can be expressed in terms of a graph-based format. The present system makes use of this by employing simple interfaces that translate syntactic annotations, which are acquired either via parsing with XLE or with Universal Dependencies, into such abstract graph representations. These can then be modified by using, in principle, simple rewrite rules.

Rewrite rules consist of a query expression that serves to identify sub-graphs in a given annotation and an expansion graph that specifies the information that is added to the input provided that the query matches. Rules are specified in terms of a fact notation where a fact consists of a mother node, an attribute/relation, and a value/daughter node. Nodes are identified via variables, while attributes/relations are arbitrary strings without white spaces.

By engineering the output appropriately, it can be directly fed into further processing steps. We show this in terms of a syntax/semantics interface and a semantic interpretation component which produces semantic representations based on Glue semantics.

In summary, we present a system for expanding and rewriting linguistic annotations that can be applied to a wide array of linguistic resources given a simple translation interface,inspired by the ideas of Ide and Bunt (2010). Previous work on the PRS contained in XLE has shown that such a system can find a wide array of creative uses and opens up new possibilities for using formal computational methods in NLP. We concretely show this by presenting a syntax/semantics interface implemented with the system presented here. Since the system is implemented in a micro-service architecture, it can be easily integrated into linguistic annotation pipelines.

## References

Bobrow, Daniel G., Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price, and Annie Zaenen. 2007. PARC's Bridge and Question Answering System. In *Proceedings of the GEAF 2007 Workshop*. 1–22.

Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman. 2017. *XLE Documentation*. Palo Alto Research Center.

Ide, Nancy and Harry Bunt. 2010. Anatomy of Annotation Schemes: Mapping to GrAF. In *Proceedings of the Fourth Linguistic Annotation Workshop*. 247–25.