Moving TIGER beyond Sentence-Level

Agnieszka Falenska, Kerstin Eckart, Jonas Kuhn
Institute for Natural Language Processing
University of Stuttgart
{falenska, eckartkn, jonas}@ims.uni-stuttgart.de

Abstract

We present TIGER 2.2-doc – a new set of annotations for the German TIGER corpus. The set moves the corpus to a document level. It includes a full mapping of sentences to documents, as well as additional sentence-level and document-level annotations. The sentence-level annotations refer to the role of a sentence in the document. They introduce structure to the TIGER documents by separating headers and meta-level information from article content. Document-level annotations recover information which has been neglected in the intermediate releases of the TIGER corpus, such as document categories and publication dates of the articles. Additionally, we introduce new document-level annotations: authors and their gender. We describe the process of corpus annotation, show statistics of the obtained data and present baseline experiments for lemmatization, part-of-speech and morphological tagging, and dependency parsing. Finally, we present two example use cases: sentence boundary detection and authorship attribution. These use cases take the data from TIGER into account and illustrate the usefulness of the new annotation layers from TIGER 2.2-doc.

Keywords: Corpus Annotation, Treebank, Document Structure

1. Introduction

The TIGER corpus (Brants et al., 2004) is an important treebank for German. It has been frequently employed over the years: for several shared tasks on multilingual dependency parsing (Buchholz and Marsi, 2006) and parsing morphologically rich languages (Seddah et al., 2013; Seddah et al., 2014); as part of the HamleDT compilation (Zeman et al., 2012); and as training data for processing pipelines for variety of tasks, including coreference and bridging resolution (Björkelund et al., 2014), theoretical linguistics (Haselbach et al., 2012), and building morphological word-embeddings (Cotterell and Schütze, 2015).

TIGER consists of roughly 900,000 words or 50,000 sentences from German newspaper articles and its focus is on token and sentence level: For the first release of the TIGER treebank, the corpus was semi-automatically annotated and corrected with part-of-speech tags and constituent structures. The following versions enhanced the corpus with respect to the number of sentences and introduced morphological and lemma annotations. Additionally, parts of the corpus were released as dependency treebanks, e.g. a dependency version of TIGER 2.2 (Seeker and Kuhn, 2012).

During the early development of the corpus parts of the information about document boundaries were lost. The current distribution 2.2 does not contain any grouping of sentences into documents, i.e., the corpus consists of one long sequence of sentences. Therefore, it is not possible to annotate this dataset with any document-level information, such as anaphora and coreference relations, or discourse trees.[1] However, as opposed to datasets where contiguous documents are not available due to technical design or copyright reasons (Faaß and Eckart, 2013; Schäfer, 2015), TIGER does contain full documents and the only information lost is a mapping between sentences and documents to which they belong.

The purpose of this paper is to overcome shortcomings of the current TIGER distribution and to enable application of document-level tasks to it. We present TIGER 2.2-doc[2], a new set of annotations that divides the sentences of the corpus into documents, i.e., newspaper articles, and realigns articles with their categories and approximate publication dates using original sources. We also semi-automatically enrich documents with structure annotations (differentiating meta-level information from article's real content), authors, and authors' genders. Additionally, we release a new split into training, development and test sets.[3]

A corpus containing gold-standard syntactic annotations and document boundaries serves as a thorough platform for further annotations and applications, such as the frequently used German treebank TüBa-D/Z (Telljohann et al., 2015). We believe that adding new annotations to TIGER and introducing TIGER 2.2-doc will make the corpus an interesting resource for NLP researchers working on a variety of tasks. The wide range of annotations – from sentence boundaries, through syntax, to authors and categories – makes it possible to answer new questions. For example, document categories can be used to examine which types of articles are the most difficult to parse. Similarly, having gold syntax trees and gender of authors allows to analyze if some syntactic constructions are used more often by women than men. In this paper we present baseline experiments on lemmatization, part-of-speech (POS) tagging, morphological analysis, and dependency parsing for the new training/development split of TIGER. We also show pilot experiments on sentence boundary detection and authorship attribution in TIGER.

[1] Cf. guidelines like Riester and Baumann (2017) and Riester et al. (to appear) which have been applied to German text, taking documents into account.

[2] Persistent identifier (PID): `http://hdl.handle.net/11022/1007-0000-0000-8E50-6`

[3] In TIGER 2.2 one additional consequence of the sentence-level focus is that the typical training, development, and test set split does not take document borders into account and contains sentences from the same documents in different sets.

2. (Re-)introducing TIGER Documents

TIGER is provided in several representation formats, the set of which differs among the releases. In the Negra Export format (Brants, 1997), some metadata was encoded in a section called ORIGIN. An entry in this section consists of a numeric identifier, which we call the **ORIGINID**, information on an approximate publication time and, partially, information on article categories (news, feuilleton, etc.). According to the dates from the metadata, the corpus can be split into articles from years 1992, 1995 and 1997.

A sentence is related to the metadata via the ORIGINID. We consider a set of consecutive sentences annotated with the same ORIGINID an article and would like our documents to contain exactly one article each. However, the information in the corpus releases was not sufficient to comprehensively identify the document boundaries: (1) For the 1992 part, the metadata shows a different granularity by subsuming all articles under one ORIGINID, so no document boundaries could be extracted from the metadata for this part. (2) Part of the mapping between sentences and ORIGINID got lost during processing, such that several blocks of sentences were assigned a fall-back ORIGINID 0, which does not contain any information. (3) Some presumed articles contain more than one or only parts of articles, maybe due to extraction errors when the articles were selected from the newspaper data.

To introduce a full sentence to document mapping we used additional resources and some old work files from the creation phase of the corpus. Since the additional information usually referred only to one of the corpus parts (1992, 1995 or 1997), we treated each part separately. However, some common guidelines were applied:

1. TIGER 2.1 sentence numbers and segments are kept;
2. ORIGINIDs are kept to align the metadata;
3. When applying additional resources with article markup their boundaries should be reflected in the annotation as long as they do not conflict with the boundaries introduced by the ORIGINIDs.

TIGER 2.2-doc thus introduces **DOCUMENTIDs** X_Y where X contains an ORIGINID and Y is an addition to mark several documents within one ORIGINID.[4]

For the 1992 part we automatically extracted article borders from ECI/MC1 (European Corpus Initiative, 1994). After minimal manual post-processing all documents were read by a native speaker who also compared the proposed boundaries to the ECI/MC1 data.

For the 1995 part we applied an old work file. The file contained a full mapping of the ORIGINIDs but different sentence numbers and sentence segmentation. For each range with ORIGINID 0 the ORIGINIDs from the work file were automatically mapped when the sentences were identical and unique. Afterwards all documents were read by native speakers. Taking the work file into account, they assigned and corrected DOCUMENTIDs where necessary.

For the 1997 part we applied an old work file and a part of the DeReKo (Institut für Deutsche Sprache, nd). Again, identical and unique sentences with ORIGINID 0 were automatically mapped to the work file. Afterwards all documents were read by native speakers. Taking the work file and DeReKo into account, they assigned and corrected documentIDs where necessary. A range of sentences with ORIGINID 0 remained, for which only the metadata information was available. The missing DOCUMENTIDs were assigned manually based on the number of ORIGINIDs left for this part and the respective article categories.

During the process of (re-)introducing document boundaries, the annotators found special cases in the corpus. Some of these can be explained as processing artifacts, such as where a sentence from one article was copied accidentally into another article, while others reflect peculiarities of the newspaper, such as collection articles, where a set of independent newsflashes is combined under one heading and consequently marked with the same ORIGINID in the corpus. Collection articles however influence document-level applications such as coreference annotations, because one document consists of several small parts which usually do not share any referents. Based on Guideline 3 mentioned above, these collection articles were handled according to their markup in the external resources and treated either as one or as several documents. Respective comments are released with the annotation of TIGER 2.2-doc to let users decide to e.g. exclude some article types.

3. Further Annotations

TIGER 2.2-doc includes a mapping from sentences to 2,263 documents, with on average 22 sentences per document. The documents are split into training/development/test sets with 1,863/200/200 documents. An example document with 21 sentences is shown in Figure 1. Clearly the document in its original published form contained an internal structure – the first sentence was a title, followed by a subtitle, meta information about the author, and place and date of release. We introduce the structure to the documents and annotate them on sentence and document level. However, we exclude from this annotation process 100 collection documents, i.e., documents consisting of more than one article, since they have their own structure.

3.1. Sentence-level Annotations

We introduce a new sentence-level layer of annotations with three possible values (see Figure 1 for examples): (i) HEADER for all sentences which act as titles (in this version we do not distinguish titles from subtitles); (ii) META for all sentences which contain meta-level document information, as author, date of release of the article, notes for the reader; (iii) BODY for the actual content of the article.

We annotated all sentences with HEADER, META, or BODY semi-automatically. First all sentences were annotated automatically with a rule-based system. The system checked if a sentence ends with a sentence-final punctuation mark (. ? ; or !), is capitalized, contains verb, or contains one of manually selected key words, e.g., "Von" (eng. *by*), "Kommentar" (eng. *(editorial) comment)*, or "Seite" (eng. *page*). Then the annotations were checked manually by a native speaker and corrected. Approximately 1% of sentences received a wrong annotation during the automatic process, in most cases HEADER was incorrectly annotated as BODY.

[4] Both parts are four digit values with leading zeros, i.e. 0001_0005 is the fifth article found under ORIGINID 1.

1	SPD-Spitze stimmt Bosnien-Einsatz zu *Head of SPD [Social Democratic Party of Germany] agrees to Bosnia mission*	HEADER
2	Parteitag soll Engagement deutscher Soldaten “ ohne Kampfauftrag ” beschließen *Party convention is to give consent on German troops engaging in "non-combat" mission*	HEADER
3	Von Helmut Lölhöffel *By Helmut Lölhöffel*	META
4	MANNHEIM , 13. November .	META
5	Die SPD ist bereit , dem Einsatz von Bundeswehr-Transport- und Aufklärungsflugzeugen bei der Umsetzung eines Bosnien-Friedensplanes zuzustimmen . *The SPD is willing to go along with having German Federal Defense Forces operate transport and reconnaissance aircraft in implementing a road map for peace for Bosnia.*	BODY
6	Damit folgt die Partei der sozialdemokratischen Bundestagsfraktion , die dies schon vorher beschlossen hatte . *By doing so, the party follows the social democratic members of German parliament, who have already agreed on this.*	BODY
...	...	
21	Leitartikel auf Seite 3 *Editorial on page 3*	META

Figure 1: An example document (0834_0001) and its translation to English (translation is not a part of the treebank). Document-level annotations for this document are: category – NAC, author – Helmut Lölhöffel, author's gender – male, approximate date of publication – 1995-11-14. The publication date may differ from the one in the article ("13. November" v. "1995-11-14") because it is approximate and refers to the week in which the article was published.

In the final version 6% of all sentences is annotated with META and 9% with HEADER.

3.2. Document-level Annotations

Document-level annotations are categories, approximate dates of publication, and authors. The first two types were recovered from the metadata of the original TIGER sources, while authors were annotated semi-automatically (see details below). The distribution of document-level annotations across all documents is presented in Figure 2. Approximate publication dates are available for all documents, categories for 2,016 of them, and authors for 541. For 506 documents all three types of annotations are available.

Categories The document categories were recovered from existing sources – see Table 1 for details. For six big categories (news, economy, world news, feuilleton, politics, and scientific topics) both the category ID and meaning of it were recovered. 1,749 documents received one of these categories. For additional 267 documents (under ”Various categories” in the Table 1) only the category ID was restored. For 247 documents category is unknown.

Category		#documents	Meaning
Nachrichten	NAC	689	*News*
Wirtschaft	WIR	601	*Economy*
Other		267	*Various categories*
–		247	*Unknown category*
Aus Aller Welt	AAW	177	*World news*
Feuilleton	FEU	148	*Feuilleton*
Politik	POL	106	*Politics*
Wissenschaft	WIS	28	*Scientific topics*

Table 1: Document categories.

Publication Dates The publication dates available in the metadata are approximate publication dates for the articles, probably referring to the week in which an article was published. By means of the ORIGINID this information is available for all documents.

Authors Information about the author of the document comes from the article's content – see for example Sentence 3 in Figure 1. We used sentence-level annotations to collect two types of meta annotations: the author of the article and the gender of the author. First we found all sentences with META annotations. Then we used gold POS tags to filter phrases with NE tags. These phrases were manually checked and the correct authors were collected. Finally we used morphological features to mark genders of the authors. For example, token level annotations for Sentence 3 from Figure 1 are: Von/`APPR/-` Helmut/`NE/case=nom|num=sg|gender=masc` Lölhöffel/`NE/case=dat|num=sg|gender=masc`. We used POS tags to filter "Helmut Lölhöffel" as a possible author and values of the gender morphological feature to set the gender of the document's author to "male".
The information about authors is available for 541 TIGER documents: 8 of the articles were written by more than one author, all the other 533 articles were written by 243 different authors, among which 87 wrote more than one article. We present the distribution of authors in Figure 3.

4. Baseline Experiments

The TIGER corpus has been widely employed to evaluate methods predicting token-level annotations, as POS tags or dependency trees (Buchholz and Marsi, 2006; Seddah et al., 2014, among others). These token-level annotations do not differ between TIGER 2.2 and TIGER 2.2-doc. However, TIGER 2.2-doc is released with a new split into training, development, and test sets. To enable future comparisons, we present baseline results for lemmatization, part-of-speech tagging, morphological analysis, and dependency parsing for the new split.

Methodology and Tools We predict POS tags and morphological features with the state-of-the-art morphological

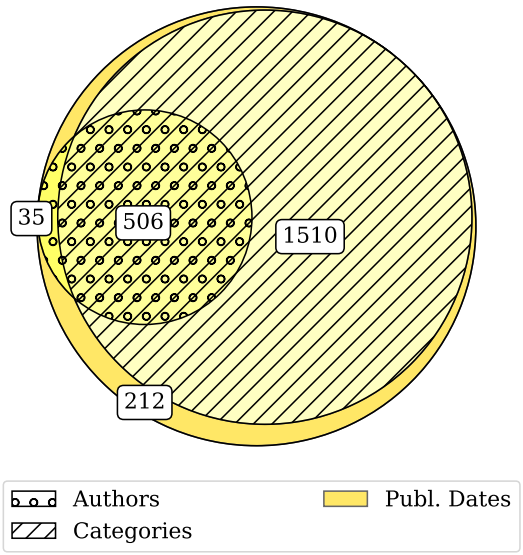


Figure 2: Number of documents annotated with different document-level annotations. Publication dates are available for all 2,263 documents. 2,016 of them (1510 + 506) are annotated with categories and 541 (506 + 35) with authors. For 506 documents all three types of annotations are available.

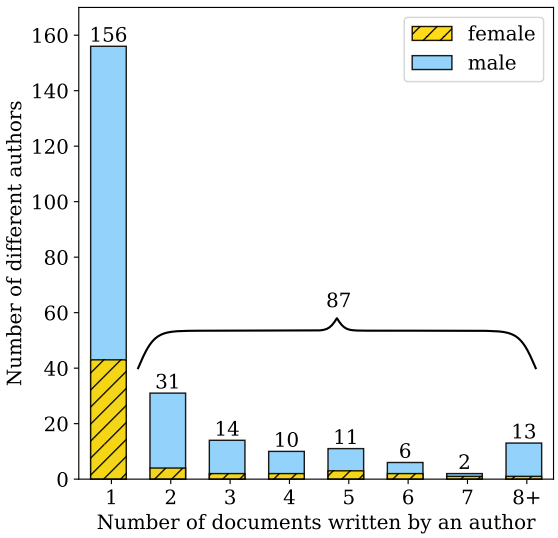


Figure 3: Distribution of authors of TIGER articles and gender of the authors. 8 articles which were written by more than one author are not included in the plot. Remaining 533 articles were written by 243 different authors, among which 87 wrote more than one article.

CRF-based tagger (Müller et al., 2013). We use the mate-tools for lemmatization.[5] For dependency parsing we apply a transition-based beam search parser by Björkelund and Nivre (2015) which uses the ArcStandard system extended with a Swap transition (Nivre, 2009). We train the parser on 10-fold jackknifed POS and morphological tags.

Results We provide results for the new development and test sets in Table 2.

	Lemma F1	POS F1	Morph F1	Parsing LAS
Dev	97.96	97.78	90.68	91.27
Test	97.91	97.86	90.63	92.53

Table 2: Baseline results for lemmatization, part-of-speech tagging, morphological analysis, and dependency parsing on the development and test sets.

5. Use cases

To illustrate the usefulness of the new annotations, we present two use cases: sentence boundary detection and authorship attribution. Both tasks take document borders into account and could not have been applied on TIGER before introducing the new annotation layers. While gold-standard sentence boundaries were available in TIGER 2.2, TIGER 2.2-doc introduces the distinction between sentence boundaries and document boundaries. The new information allows to train and evaluate sentence segmenters only on meaningful instances of sentence borders, and not the ones at the end of documents.

[5] https://code.google.com/archive/p/mate-tools/

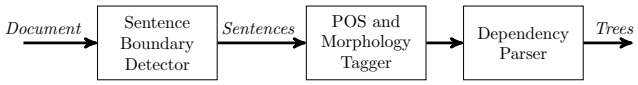
5.1. Case Study: Sentence Boundary Detection

The logical structure of a document describes of what units (e.g., titles, enumerations, tables) the document consists. Such information is useful for automatic document processing tasks, such as information extraction or summarization. Document structure reconstruction is the task of automatically recovering the logical structure of documents, for example from the results of an OCR process.

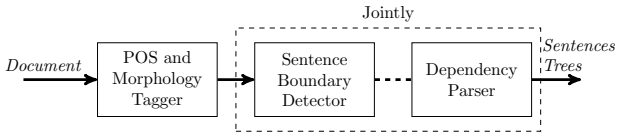
The original newspaper versions of TIGER articles were structured and contained visually distinguishable titles, subtitles, author's footnotes, etc. In the current version some of this information was reconstructed and marked with sentence level annotations, making TIGER 2.2-doc a possible experimental field for document structure reconstruction. In this paper we present pilot experiments for the first step of automatic document structure reconstruction: sentence boundary detection. We leave the next steps (e.g., prediction of sentence-level annotations) for future work.

Methodology Sentence boundary detection is often regarded as a solved task, at least in the domain of well-edited texts. Therefore, typical sentence boundary detectors focus on orthographic clues, as punctuation and capitalization marks. However, when moving to non-standard texts basic assumptions about punctuation and capitalization may be violated, thus rendering sentence segmentation more challenging task (Evang et al., 2013).

We have previously shown (Björkelund et al., 2016) that for texts where typical orthographic features are not present better sentence segmentation can be achieved by re-ordering the standard NLP pipeline (e.g., pipeline that applies lemmatization, part-of-speech tagging, and/or parsing). In the standard approach, sentence segmentation is regarded as the easiest task and is performed first (see Figure 4a). Alternatively, we trained a POS tagger on whole documents, applied it as the first tool in the pipeline, and



(a) standard pipeline



(b) re-ordered pipeline

Figure 4: Schematic view of pipelines applied in the experiments.

	Segm. F1	POS F1	Morph F1	Parsing LAS
Standard pipeline				
UDPIPE	82.34	97.66	90.45	87.42
MARMOT	89.21	97.74	90.58	88.47
Re-ordered pipeline				
JOINT	93.26	97.73	90.19	88.53
JOINT-REPARSED				88.92

Table 3: Results on the development set for two pipelines and three sentence segmenters.

performed sentence segmentation and parsing jointly (see Figure 4b). We showed that syntax is useful for sentence segmentation and that the joint method detects sentence boundaries better.

TIGER is built from newspaper articles which are usually considered well-edited texts and common sentence segmenters should perform well for them. However, TIGER documents also contain meta-level sentences and headers which are not straightforward to segment. We perform pilot experiments for sentence segmentation in TIGER, apply state-of-the art tools and investigate if the joint method of Björkelund et al. (2016) is a better choice for this dataset.

Tools We apply state-of-the-art sentence segmenters to establish baseline results:

UDPIPE (Straka et al., 2016): is a toolset performing, i.a., sentence segmentation, POS tagging, and dependency parsing. From all the functionalities of UDPIPE we employ only the sentence segmenter. It uses a single-layer bidirectional GRU network. The segmenter works on character level, i.e., for each character in text it predicts if a sentence ends after it. We use UDPIPE as a baseline following the recent CoNLL Shared Task 2017 (Zeman et al., 2017).

MARMOT: is the best performing baseline from Björkelund et al. (2016). It augments POS tags with information if a token starts a new sentence. Then, it trains the sequence labeler of Müller et al. (2013) on whole documents annotated with the augmented tags. The method applied to new documents jointly predicts sentence boundaries and POS tags.

For POS tagging, morphological analysis and parsing we apply the methods from Section 4 – namely CRF-based tagger by Müller et al. (2013) and transition-based parser by Björkelund and Nivre (2015).

For the re-ordered pipeline we train the same POS tagger but on whole documents instead of sentences. We employ the aforementioned parser extended to predict sentence boundaries (Björkelund et al., 2016) (referred to as **JOINT**). It predicts sentence boundaries and dependency trees jointly. For both pipelines we train parsers on 10-fold jackknifed POS and morphological tags. We calculate the results with the evaluation script from the CoNLL Shared Task 2017 (Zeman et al., 2017).

Results The results of the experiments are presented in Table 3 in the first column. We find that MARMOT outperforms the UDPIPE baseline by a big margin of almost 7 points. JOINT surpasses both baselines – by 4 points for MARMOT and almost 11 points for UDPIPE, making it a better choice for the first step of document structure reconstruction in the TIGER dataset.[6]

We also investigate how the next pipeline steps are influenced by different sentence segmentation results. POS tagging and morphological analysis are almost not influenced by the selection of the pipeline, i.e., regardless if they are applied to sentences or documents their results are very similar. When comparing parsing results we see that sentence boundary detection errors propagate through the pipeline. The final parser loses one point when applied to sentences predicted by UDPIPE instead of MARMOT. Interestingly, syntactic features in the JOINT method help the sentence boundary detection but not the other way around, i.e., its parsing result outperforms both baselines only slightly. To assess the influence of error propagation through the pipeline we follow Björkelund et al. (2016) and parse the sentences once again (denoted **JOINT-REPARSED**). In this scenario JOINT serves only as a sentence segmenter and parsing is performed separately. We find that the better sentence segmentation translates to better parsing results and JOINT-REPARSED outperforms both MARMOT and UDPIPE.

5.2. Case Study: Authorship Attribution

Authorship attribution is the task of determining the author of a document. TIGER is built from short newspaper texts which mostly belong to categories related to news. It is an interesting question if it is possible to track documents' authors in this domain. And if yes, if it is due to stylometric differences or content of the documents (e.g. authors tending to write documents on the same topics). We present pilot experiments for two authorship attributions tasks: predicting the author of the text and gender of the author.

Methodology and Tools In both of the tasks we use only sentences not labeled as META, as they contain implicit information on authors (names, surnames and cities). For gender prediction we use all 533 documents with author id (for a detailed distribution of documents and their authors

[6]We also applied the two described pipelines to an out-of-domain test suite (Seeker and Kuhn, 2014). Results showed that JOINT trained on the new TIGER 2.2-doc has an advantage over standard pipeline also for other diverse datasets – for example dvd player manuals.

see Figure 3). For author attribution we use only authors who wrote at least two documents (87 authors and 377 documents). We use two classification methods:

DELTA: the Burrows's Delta (Burrows, 2002) is the most established method for capturing stylometric differences. We calculate DELTA and perform classification with the R stylo package (Eder et al., 2013) with default parameters. To establish if the method captures stylometric or content differences we follow Schulz et al. (2016) and calculate DELTA in two ways: on all the words in the documents (DELTA–STYLE) and on content words (DELTA–CONTENT). As content words we select nouns, verbs and adjectives and filter them by gold-standard POS tags.

VERETAL: we copy the experimental setup from Verhoeven et al. (2016). We use LinearSVC from sklearn[7] with default parameters and use unigrams and bigrams of words and trigrams and tetragrams of characters as features.

We compare the two methods to two control baselines: weighted random baseline (**WRB**), which predicts a stratified random class, and majority baseline (**MAJ**), which predicts the most frequent class. We evaluate the accuracy with 10-fold cross-validation.

Results Table 4 presents results of the experiments. In the task of gender prediction documents are assigned only two classes and the methods achieve higher accuracy. Among all the methods, WRB is the worst and is able to correctly predict authors' gender only for 69.03% of the documents. DELTA slightly outperforms WRB, especially when using content words. This might be an indicator that the "author signal" in the TIGER documents is weak and classification methods capture more content than stylometric differences. Generally, DELTA is not able to outperform the simple MAJ baseline, which achieves accuracy of 79.91%. But VERETAL, a classifier with a richer set of features, proves to be the best method and outperforms MAJ by a margin of 2.43 points.

Results for author prediction are generally lower than for gender prediction because the task is harder – it involves assigning one of 87 classes. In this case DELTA is able to outperform both control baselines. Interestingly, this time style features give a small boost over the content ones. Again, VERETAL surpasses all the other methods with a big margin of more than 32 points.

	Author prediction	Gender prediction
WRB	2.71	69.03
MAJ	2.99	79.91
DELTA–STYLE	12.16	71.08
DELTA–CONTENT	9.75	72.97
VERETAL	44.41	82.34

Table 4: Accuracy (10-fold cross-validated) for predicting author (87 authors, 377 documents) and gender (female/male, 533 documents).

[7] scikit-learn.org/

6. Conclusion

We presented TIGER 2.2-doc, a new set of annotations for the frequently used German corpus TIGER. Prior releases of TIGER already contained token and sentence boundaries. TIGER 2.2-doc adds explicit document borders based on the newspaper articles, and a suggested split into training, development and test sets. We presented new document- and sentence-level annotations which broaden the range of possible applications of TIGER. We showed two example use cases which employ the new annotations: sentence boundary detection (as a part of document structure reconstruction) and authorship attribution. TIGER 2.2-doc is available by means of a persistent identifier.

7. Acknowledgements

We would like to thank Amir Zeldes and Florian Zipser for discussions on the (re-)introduction of the document borders and for help with finding the old work files, and our student annotators for patiently mapping and reading lots of old news. This work was supported by the German Federal Ministry of Education and Research (BMBF) via CLARIN-D and the German Research Foundation (DFG) via SFB 732, projects D8 and INF.

8. Bibliographical References

Björkelund, A. and Nivre, J. (2015). Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 76–86, Bilbao, Spain. Association for Computational Linguistics.

Björkelund, A., Eckart, K., Riester, A., Schauffler, N., and Schweitzer, K. (2014). The extended DIRNDL corpus as a resource for coreference and bridging resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3222–3228, Reykjavik, Iceland. European Language Resources Association (ELRA).

Björkelund, A., Faleńska, A., Seeker, W., and Kuhn, J. (2016). How to train dependency parsers with inexact search for joint sentence boundary detection and parsing of entire documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1934, Berlin, Germany. Association for Computational Linguistics.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Brants, T. (1997). The NEGRA export format. CLAUS Report 98, Universität des Saarlandes, Computerlinguistik, Saarbrücken, Germany.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Burrows, J. (2002). âĂŸDeltaâĂŹ: a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.

Cotterell, R. and Schütze, H. (2015). Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Eder, M., Kestemont, M., and Rybicki, J. (2013). Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*, pages 487–489. University of Nebraska, Lincoln Lincoln, NE.

Evang, K., Basile, V., Chrupała, G., and Bos, J. (2013). Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426. Association for Computational Linguistics.

Faaß, G. and Eckart, K. (2013). SdeWaC – a corpus of parsable sentences from the web. In Iryna Gurevych, et al., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

Haselbach, B., Eckart, K., Seeker, W., Eberle, K., and Heid, U. (2012). Approximating theoretical linguistics classification in real data: the case of German "nach" particle verbs. In *Proceedings of COLING 2012*, pages 1113–1128. The COLING 2012 Organizing Committee.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 351–359, Suntec, Singapore. Association for Computational Linguistics.

Riester, A. and Baumann, S. (2017). *The RefLex Scheme – Annotation Guidelines*, volume 14 of *SinSpeC. Working Papers of the SFB 732*. University of Stuttgart.

Riester, A., Brunetti, L., and Kuthy, K. D. (to appear). Annotation guidelines for questions under discussion and information structure. In Katharina Haude Evangelia Adamou et al., editors, *Information Structure in Lesser-Described Languages*. Benjamins, Amsterdam.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, et al., editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

Schulz, S., Kuhn, J., and Reiter, N. (2016). Authorship attribution of mediaeval German text: Style and contents in Apollonius von Tyrland. In *Digital Humanities 2016: Conference Abstracts*, pages 883–885, KrakÃşw, Poland.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., and de la Clergerie, E. V. (2013). Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages.

Seddah, D., Kübler, S., and Tsarfaty, R. (2014). Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages.

Seeker, W. and Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a German treebank. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3132–3139, Istanbul, Turkey. European Language Resources Association (ELRA).

Seeker, W. and Kuhn, J. (2014). An out-of-domain test suite for dependency parsing of German. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4066–4073, Reykjavik, Iceland. European Language Resources Association (ELRA).

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, PortoroÅ¿, Slovenia. European Language Resources Association (ELRA).

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., and Beck, K. (2015). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar fur Sprachwissenschaft, Universität Tübingen, Tübingen, Germany*.

Verhoeven, B., Daelemans, W., and Plank, B. (2016). TwiSty: A multilingual twitter stylometry corpus for gender and personality profiling. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1632–1637, PortoroÅ¿, Slovenia. European Language Resources Association (ELRA).

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). HamleDT: To Parse or Not to Parse? In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2735–2741, Istanbul, Turkey. European Language Resources Association (ELRA).

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Mis-

silä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

9. Language Resource References

European Corpus Initiative. (1994). *European Corpus Initiative Multilingual Corpus 1 (ECI/MC1)*. Linguistic Data Consortium, ELRA, 1.0, ISLRN 511-168-567-582-5.

Institut für Deutsche Sprache. (n.d.). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache*. Institut für Deutsche Sprache.