NoSta-D: A Corpus of German Non-Standard Varieties

Stefanie Dipper[1], **Anke Lüdeling**[2], **Marc Reznicek**[1,2]
Ruhr-Universität Bochum[1]
Humboldt-Universität zu Berlin[2]

Abstract

Until recently, most research in computational linguistics has been done on newspaper texts. Nowadays, the focus has been extended to other types of language data. This means that many linguistic descriptions and automatic tools need to be adapted or extended to non-newspaper language. The non-standard varieties corpus of German (*NoSta-D*) will provide a first gold standard for evaluation and training data of dependency analysis, named entity recognition and coreference resolution for *out-of-domain* text types.

1 Introduction[1]

Even though the first electronically available corpus was a historical one [2], it can be said that in computational linguistics most corpora mainly consist of standard written text such as newspaper texts.[2] This is due to the fact that newspaper text (or other standard written texts such as technical manuals) can be accessed without any problem via the internet, and

[1]The research reported here has been financed by the German Federal Ministry of Education and Research (BMBF). All links were checked on Dec 12th, 2012.

[2]This is true even though there are a number of corpora that contain texts from different varieties (such as many of the large reference corpora) and there are now many historical corpora, dialect corpora, or corpora of other varieties.

often adhere to a format that can be processed rather easily. As a consequence, tool and schema development in computational linguistics has tended to focus on standard written texts and, due to that bias, current-state annotation tools and guidelines are tuned towards newspaper texts, which in turn has become the *de facto* "standard variety".[3]

A growing pool of studies on different text types and varieties (e.g. chat and blog data from the internet, learner data, historical texts) demonstrate the limits of the current systems.[4] Many linguistic structures occurring in these "non-standard varieties" are not covered by the tag sets and annotation schemes currently in use.

In this short paper, we present a pilot corpus of non-standard varieties for German, *NoSta-D*, which is compiled as part of the CLARIN-D curation project 'Linguistic Annotation of Non-standard Varieties — Guidelines and Best Practices' at Humboldt-University Berlin and Ruhr-University Bochum. *NoSta-D* consists of (small) subcorpora of different varieties (see Section 2) that are being annotated with dependency syntax, named entities, and coreference using schemes and tools originally developed for newspaper text (see Section 3). The corpus is used for two purposes: First, by annotating the corpus with the same schemes (rather than developing different schemes for each variety) we can describe and quantify the differences between the varieties. And second, these findings can then be used to identify shortcomings of current guidelines and tools. One of the project results will be the *NoSta-D* corpus including gold standard annotations at all three annotation levels. *NoSta-D* will be made freely available.[5]

[3]The term 'standard' is not intended as a normative concept, but refers to the *de facto* standard language found in newspaper texts.

[4]This is outside the scope of this paper but it can be said that many of the available corpora of the 'nonstandard' varieties are collected in linguistics proper rather than in computational linguistics — many of them are not (deeply) annotated and not freely available. This is slowly changing, as the knowledge of corpus architecture, formats, and tools is becoming more and more available in linguistics and as the focus in linguistics is shifting from standard written language towards the study of variation. For more on these issues, see, e.g., the papers in [11].

[5]At the CLARIN-D project page in September 2013, see
`http://clarin-d.de/en/discipline-specific-working-groups/wg-7-applied-linguistics-computational-linguistics/curation-project-2`.

2 The Corpus

NoSta-D consists of texts from five non-standard varieties which are selected with the aim to cover a broad range of linguistic variation and non-standard phenomena: historical data, chat data, spoken data, learner data, and literary prose, see the overview in Table 1. All subcorpora stem from already existing research projects.[6] In addition, a part of the newspaper corpus TüBa-D/Z has been included to provide a baseline for annotation evaluations. We only include subcorpora that are free of copyright restrictions.[7]

	Subcorpus	Variety	# Tokens	Provider
1	DDB Anselm Corpus	historical	2,348 4,705	Berlin Bochum
2	Dortmunder Chat Corpus	chat	6,664	Dortmund
3	BeMaTaC	spoken	6,731	Berlin
4	Falko	learner	6,762	Berlin
5	Kafka: Der Prozeß	literary prose	7,294	DigBib.Org
6	Tüba-D/Z (subset)	newspaper	5,000	Tübingen

Table 1: *NoSta-D* corpus design: the subcorpora

As the corpus has to be annotated manually, within a limited amount of time, the amount of text for each variety that will be annotated in the first round is quite small (~ 300 sentences or utterances, ~ 7,000 tokens). Careful selection assures that the included passages and texts show a high rate of interesting linguistic structures.

[6]The subcorpora come from the following projects: DDB: `http://korpling.german.hu-berlin.de/ddb-doku/`, Anselm: `http://www.linguistics.ruhr-uni-bochum.de/anselm`; Dortmunder Chat Corpus: `http://www.chatkorpus.tu-dortmund.de/korpora.html`; BeMaTaC: `http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematac`; Falko: `http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko`; Kafka: Der Prozeß: `http://www.digbib.org/Franz_Kafka_1883/Der_Prozess`. The literary-prose corpus is new and covers the growing demand in eHumanities.

[7]For licencing TüBa-D/Z, see `http://www.sfs.uni-tuebingen.de/resources/tuebadz-license.html`

Formats In conformity with the proposals of the ISO Technical group TC37/SC4[8], all data are stored in stand-off formats to allow for unrestricted later addition of annotations. As part of the curation project, converters are being developed to assure interchangeability between the TCF format used in the CLARIN-D webservice environment WebLicht [4] and the manual annotation tool WebAnno[9] on the one hand, and more generic corpus formats such as PAULA [3] for storage and relAnnis for the corpus search tool ANNIS [12] on the other hand.

3 Preprocessing and Annotations

Tokenization The first processing step consists of marking word and sentence boundaries. Current tokenizers usually cannot deal with many spelling phenomena of non-standard data. Chat data, e.g., contains emoticons and other types of special symbols in various forms, see Ex. (1). Sentence boundary detection is especially difficult with historical data, which sometimes does not use punctuation marks at all, and sometimes uses punctuations for purposes that differ from modern use, such as marking prosodic or phrase boundaries, see Ex. (2). Furthermore, word boundaries in historical data also diverge from modern boundaries. For instance, *wiltu* in Ex. (2) corresponds to the modern sequence *willst du* 'want you'.

(1) winke@bochum :-))
'wave to Bochum :-))'

(2) Wiltu nu gvter menſche· eynen guten bowm ſeen vnd· wiltu gute frucht an dyner zele brengen· ſo ſalt u dich vben an guten werken·

'If you good human want to seed· a good tree and· if you want to bring good fruit to your soul· then you should exercise in good deeds·'

Normalization Data used as training or evaluation data in computational linguistics must be consistently annotated. Hence, finding ways to assure

[8] http://www.tc37sc4.org

[9] The annotation tool WebAnno is developed as part of the CLARIN-D curation project 'Implementation of a web-based platform for linguistic annotations', http://www.ukp.tu-darmstadt.de/research/current-projects/clarin-d/.

consistent decisions for *variable* and *non-standard* data is an important issue, see [1].

Ex. (3) shows an example from the chat data. In standard German, the preposition *mit* 'with' selects dative case. In the chat data, *mit* seems to occur with accusative case.

(3)	**Chat**	ich	versteh	mich	mit	jeden[acc]	man
	Std	Ich	verstehe	mich	mit	jedem[dat]	Mann
		'I get on with any man'					

How should a grammatical analysis (a human annotator, parser etc.) deal with such mismatches? This depends on the research question. An obvious way to deal with such "errors" would be to relax the condition on case. For instance, annotators would be told to annotate the NP following the preposition *mit* always as the complement of the preposition, regardless of the NP's case.

The relaxation of grammar constraints may lead to a more robust parser but there are two problems: If grammar restrictions are relaxed in order to deal with mismatches between the standard grammar and the variety, the information about the specific properties of that variety (with regard to a given 'standard') is not encoded explicitly. And second, because every relaxation of a grammar rule involves interpretation of the data, there are often multiple ways of 'fixing' a mismatch (this has been shown time and again for learner language, see [6]).

Consider Ex. (4), which can be fixed in (at least) two ways: In option Std1, *dass* is considered an orthographic variant of the article *das* 'the', thus providing the obligatory determiner of the count noun *examen* 'exam' (which is missing). Furthermore, the word order is marked: in the unmarked order, the adverb *morgen* 'tomorrow' would follow the verb *kommen* 'come'. In the alternative option Std2, *dass* is considered a subordinate conjunction, 'that'. Then the last three tokens would have to be switched and the obligatory article would be missing. Elements standardized as described have been put in italics in the example.

(4)	**Learner**	Ich	denke,	dass	examen	soll	kommen	morgen
	Gloss	I	think	that	exam	should	tomorrow	come
	Std1	Ich	denke,	*das*	Examen	soll	*morgen*	*kommen*
	Std2	Ich	denke,	dass *das*	Examen	*morgen*	*kommen*	*soll*

It is obvious that simple relaxation of grammar rules is not an option here. We have therefore chosen to explicitly state an interpretation of the utterances that can be dealt with by standard grammar. In this way we can (a) precisely describe the differences between the varieties, and (b) see exactly at which points the schemas and tools need to be changed to deal with each variety (rather than assume a cover-all rule relaxation). We call our interpretation of the data 'normalization', which we use as a techical term with no further theoretical implications. In some cases, we will include different normalizations to make ambiguity visible (for competing target hypotheses in learner language, see [8]).

Annotations Where necessary, data will be normalized before further annotation takes place. Next, the data will be automatically POS-tagged and lemmatized (applying the TreeTagger [9] and RFTagger [10]) and manually corrected.

For further annotation levels, we selected levels that (i) represent core tasks of computational linguistics, (ii) would provide us with interesting non-standard phenomena, and (iii) illustrate different data structures of annotation: sentence-internal pointer relations for dependency annotations, span-based annotations for named entities, and cross-sentential pointers for coreference annotations.

Dependency relations: Comparative studies on syntactic annotations have shown that languages with relatively free word order, such as German, can be described more accurately with dependency relations than with constituent structures, since dependency relations do not rely on adjacency (e.g. [7], [5]). This should make them suitable for the annotation of data from non-standard varieties, such as Ex. (4). One of the goals of our project is to investigate to what extent dependency theory is able to deal with the broad range of variation that we observe in *NoSta-D*.

Named entities: In chat data, people often use '@' to address other users, which facilitates identification of person names. In Ex. (1) above, the speaker uses a city's name (Bochum), in non-standard lower case, to refer to a user from this location. Hence, the substring *bochum* in Ex. (1) should be annotated with PER(son) rather than LOC(ation).

Coreference: In some varieties of non-standard language, coreference annotation faces special problems. For instance, certain topic constituents in spoken language can be dropped, such as the anaphoric pronoun in B's contribution (added in parentheses), which co-refers with the constituent in italics in A's contribution in Ex. (5).

(5) **A:** und dann ziehst du die Linie einmal über das gesamte Blatt bis du oben an der Ecke *der Ähre oder irgendwie Weizen oder was auch immer das da is*
B: okay (*das*) hab ich nich

A: 'and then you draw the line once over the whole sheet until you arrive on top at the corner of *the spike or wheat or whatever that is*'
B: 'okay (*this*) I don't have'

To sum up, we have presented *NoSta-D*, a corpus of German non-standard varieties. The goal of our project is (i) to come up with a (pilot) reference corpus of non-standard data, (ii) to test the coverage of current annotation guidelines or linguistic descriptions, and (iii) to evaluate the performance of state-of-the-art tools. Based on our experience with these tasks, we will come up with extended guidelines and "best practices" as to how to deal with such data.

References

[1] J. Balsa and G. Lopes. A distributed approach for a robust and evolving NLP system. In *Proceedings of NLP 2000*, pages 151–161, 2000.

[2] R. Busa. The annals of humanities computing: The Index Thomisticus. *Computers and the Humanities*, 14:83–90, 1980.

[3] C. Chiarcos, S. Dipper, M. Götze, J. Ritz, and M. Stede. A flexible framework for integrating annotations from different tools and tagsets. In *Proceeding of the Conference on Global Interoperability for Language Resources*, 2008.

[4] M. Hinrichs, T. Zastrow, and E. W. Hinrichs. WebLicht: Web-based LRT services in a distributed eScience infrastructure. In *Proceedings of LREC 2010*, pages 489–493, 2010.

[5] S. Kübler and J. Prokic. Why is German dependency parsing more reliable than constituent parsing? In *Proceedings of TLT 5*, pages 7–18, 2006.

[6] A. Lüdeling, S. Doolittle, H. Hirschmann, K. Schmidt, and M. Walter. Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45(2):67–73, 2008.

[7] J. Nivre, J. Nilsson, J. Hall, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(1):1–41, 2007.

[8] M. Reznicek, A. Lüdeling, and H. Hirschmann. Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In A. Díaz-Negrillo, editor, *Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins, to appear.

[9] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[10] H. Schmid and F. Laws. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, pages 777–784, 2008.

[11] F. Seifart, G. Haig, N. P. Himmelmann, D. Jung, A. Margetts, and P. Trilsbeek, editors. *Potentials of Language Documentation: Methods, Analyses, and Utilization*, volume 3 of *Language Documentation & Conservation*. University of Hawai'i Press, 2012.

[12] A. Zeldes, J. Ritz, A. Lüdeling, and C. Chiarcos. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, 2009.