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

Stefanie Dipper[1], Anke Lüdeling[2], Marc Reznicek[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

Newspaper texts have been among the first texts that were available electronically. Even nowadays, they represent texts that can be accessed without any problem via the internet, and often adhere to a format that can be processed rather easily. As a consequence, investigations in corpus and computational linguistics have focussed mainly on newspaper texts. Only recently have research interests been extended to other text types. 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".

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. Many linguistic structures occurring in these "non-standard varieties"[1] 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*. It has been 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. This data will be used to identify shortcomings of current guidelines and tools for dependency analysis, named entities and coreference annotations. One of the project results will be the corpus including gold standard annotations at all three annotation levels. It will be made freely available.

The paper is structured as follows. In Section 2, we present the subcorpora that our corpus consists of. In Section 3, we address preprocessing steps and the different kinds of annotations that will be added to the corpus. Section 4 conludes the paper.

2 The Corpus

We carefully selected five non-standard varieties with the aim as 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.[2] In addition, a subset of the newspaper corpus TüBa-D/Z has been included to provide a baseline for further annotation evaluations. We only included subcorpora that were free of copyright restrictions.[3].

As the corpus has to be manually annotated, within a limited amount of time, the amount of text for each variety that will be annotated in the

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

[2] The literary-prose corpus is new but covers the growing demand in eHumanities.

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

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.

xxx hier sollten wir zu allen Subkorpora Refs haben! xxx

	Subcorpus	Variety	# Tokens	Provider
1	DDB	historical	2,348	Berlin
	Anselm Corpus	historical	4,705	Bochum
2	Dortmunder Chat Corpus	chat	6,664	Dortmund
3	Bematac	spoken	~7,000	Berlin[3]
4	Falko	learner	6,762	Berlin
5	Kafka: Der Prozeß	literary prose	7,294	Tübingen
6	Tüba-D/Z	newspaper	~7,000	Tübingen

Table 1: NoSta-D corpus design: the subcorpora

2.1 Formats

In conformity with the proposals of the ISO Technical group TC37/SC4[4], 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 (Hinrichs et al. [4]) and the manual annotation tool WebAnno (xx hier eine Ref? xxx) on the one hand, and more generic corpus formats such as PAULA (Chiarcos et al. [2] for storage and relAnnis for the corpus search tool ANNIS (Zeldes et al. [14]) on the other hand.

[4] `http://www.tc37sc4.org`

3 Preprocessing and Annotations

3.1 Tokenization

The very first processing step consists of marking word and sentence boundaries. Current tokenizers usually cannot deal with many spelling phenomena of non-standard data. For instance, chat data contain emoticons and other types of special symbols in various forms, see Ex. (1). Sentence boundary detection is especially difficult with historical data. They either do not use any punctuation mark, or else they use punctuations to indicate some kind of phrase boundary, see Ex. (2). Furthermore, word boundaries in historical data also diverge from modern boundaries. For instance, *wiltu* corresponds to the modern sequence *willst du* ‘want you’.

(1) a. LANTOOO :)))

b. tag quaki :)

c. *mal guck wo quaki sich nu hinstelt*G*

(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 see· a good tree and· if you want to bring good fruit to your soul· then you should exercise in good deeds·’

3.2 Normalization

Data used as training or evaluation data in computational linguistics must be consistently annotated. Hence, finding ways to assure consistent decisions for *inconsistent* data is an important issue (see Balsa et al. [1]).

Ex (3) shows an example from the chat data. In standardized German, the preposition *mit* ‘with’ selects dative case. In the chat data, *mit* occurs with accusative case (of course, such typos can also occur in newspaper texts). The example shows the original chat data (Chat), along with a normalized version (Norm) and an English gloss (Gloss).

(3)	**Chat**	ich	versteh	mich	mit	jeden[acc]	man
	Norm	Ich	verstehe	mich	mit	jedem[dat]	Mann
	Gloss	I	understand	myself	with	any	man

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 object of the preposition, regardless of the NP’s case.

In many cases, though, relaxation of grammatical restrictions is not a solution because this can lead to multiple competing interpretations. This issue is known from learner corpus research (see Lüdeling et al. [6]). Consider Ex. (4). This sentence, which is ungrammatical, can be fixed in two ways: In option Norm1, *dass* is considered an orthographic variant of *das*, thus providing the obligatory article of the count noun *examen* ‘exam’ (which is missing). In this case, however, the word order is marked: in the unmarked order, the adverb *morgen* ‘tomorrow’ would follow the verb *kommen* ‘come’. In the alternative option Norm2, *dass* is considered a subordinate conjunction. Then the last three tokens would have to be switched and the obligatory article would be missing. Elements normalized 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
	Norm1	Ich	denke,	*das*	Examen	soll	*morgen*	*kommen*
	Norm2	Ich	denke,	dass *das*	Examen	*morgen*	*kommen*	*soll*

Obviously, the choice of normalization makes an important difference for dependency parsing.

Providing just one *robust* interpretation of such data counteracts all efforts in variationist linguistics to contrast those different expressions. In such cases we will therefore include different normalizations to make ambiguity visible (for competing target hypotheses in learner language see Reznicek et al. [11]).

3.3 Annotations

Where necessary, data will be normalized before further annotation takes place. Next, the data will be automatically POS-tagged and lemmatized (applying TreeTagger [12] and RFTagger [13]) 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 can be described more accurately with dependency relations than with constituent structures (e.g. Nivre et al. [8]). This holds for German as well (e.g. Kübler & Prokic [5]). Dependency relations can deal more flexibly with word-order variation than constituent structures since they do not rely on adjacency. 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.

Coreference In certain varieties of non-standard language, coreference annotation faces special problems. For instance, certain topic constituents in spoken language can be dropped, cf. Ex. (5). xxx hier ein echtes Bsp? xxx

(5) A: Warst du auf dem Konzert von Lu Hang?
B: Kenn ich nicht.
A: ‘Have you been to the concert by Lu Hang?’
B: ‘(That guy) I don’t know.’

Named Entities Finally, we annotate named entities as an instance of span-based annotation. We chose this xxxx hier fiel mir nichts ein. Sind chat-Daten speziell hier? z.B. weil die Teilnehmer auch immer in der 3.Person genannt werden vom Chat-System? ('Emon betritt den Raum')

3.4 Querying in ANNIS3

As has been mentioned in Section 2.1, the corpus will be searchable via the corpus search tool ANNIS. In its third version[5], a range of features that are characteristic of non-standard varieties, will be covered, e.g. overlapping token layers of multiple speakers in spoken data, or competing tokenizations (e.g. historical vs. modern word boundaries in historical data). The screenshot in **??** shall give an outlook of the corpus in its final version.

Hier ein Falko-Beispiel mit allen Annotationen bauen

4 Conclusion

In this paper, 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

[5] ANNIS3 will be released in short: http://www.sfb632.uni-potsdam.de/d1/annis/

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 the experiences that we will make during this enterprise, we will come up with extended guidelines and “best practices” as to how to deal with such data.

References

[1] Balsa, J.; Lopes, G. (2000) *A Distributed Approach for a Robust and Evolving NLP System*. Christodoulakis, D.N. (Ed.): Proceedings of the 2nd International Conference on Natural Language Processing (NLP 2000). Patras, Greece. Berlin [u.a.]: Springer, 151–161.

[2] Chiarcos, C.; Dipper, S.; Götze, M.; Ritz, J.; Stede, M. (2008) *A Flexible Framework for Integrating Annotations from Different Tools and Tagsets*. Proceeding of the Conference on Global Interoperability for Language Resources, Hong Kong.

[3] Giesel,L.; Klapi,M; Krüger,D.;Nunberger,I.;Rasskazova,O.; Sauer,S. (to appear) *Berlin Map Task Corpus (BeMaTaC):Aufbau eines L1-Vergleichskorpus zur Untersuchung gesprochener Lernersprache*. Poster to be published at CL-Postersession, DFGS 2013, Potsdam.

[4] Hinrichs,M.;Zastrow,T.; Hinrichs,E. (2011) *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10), Valetta, Malta:ELRA.

[5] Kübler, S.; Prokic, J. (2006) *Why is German Dependency Parsing more Reliable than Constituent Parsing?* Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories (TLT 2006). Prague, Czech Republic.

[6] Lüdeling, A.; Doolittle, S.; Hirschmann, H.; Schmidt, K.; Walter, M. (2008) *Das Lernerkorpus Falko*. DAF 45 (2), 67–73.

[7] Mitkov, Ruslan (2008) *Corpora for Anaphora and Coreference Resolution*. Lüdeling, A.;Kytö, M. (Eds.): Corpus Linguistics: An international Handbook. Berlin; New York: Mouton de Gruyter (Handbooks of Linguistics and Communication Science, 29,1), 579–598.

[8] Nivre, J.; Nilsson, J.; Hall, J.; Chanev, A.; Eryigit, G.; Kübler, S. (2007) *MaltParser: A Language-Independent System for Data-Driven Dependency Parsing* Natural Language Engineering 13 (1), 1–41.

[9] Rayson, P.; Stevenson, M. (2008) *Sense and Semantic Tagging*. Lüdeling, A.;Kytö, M. (Eds.): Corpus Linguistics: An international Handbook. Berlin; New York:Mouton de Gruyter (Handbooks of Linguistics and Communication Science, 29,1), 564–579.

[10] Reznicek, M.; Lüdeling, A.; Schwantuschke, F. (2012) *Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.0.* Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschu ng/falko/Falko-Handbuch_Korpusauf-bau%20und%20Annotationen_v2.01.pdf, 12-12-12.

[11] Reznicek, M.; Lüdeling, A.; Hirschmann, H. (to appear) *Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture*. DÃÂaz-Negrillo, A. (Ed.): Automatic Treatment and Analysis of Learner Corpus Data: John Benjamins.

[12] Schmid, H. (1994) *Probabilistic Part-of-Speech Tagging: Using Decision Trees*. Proceedings of the International Conference on New Methods in Language Processing, 44–49.

[13] Schmid, H.; Laws, F. (2008) *Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging.* Donia Scott (Ed.): 22nd International Conference on Computational Linguistics. Coling 2008, Stroudsburg, Pa: Association for Computational Linguistics, 777–784.

[14] Zeldes, A.; Ritz, J.; Lüdeling, A.; Chiarcos, C. (2009) *ANNIS: A Search Tool for Multi-layer Annotated Corpora*. Mahlberg,M; Gonzxxxlez-Dxxxaz, V.; Smith,C. (Eds.): Proceedings of Corpus Linguistics. Liverpool:University of Liverpool.

[15] Zipser F. (2009) *Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells*. Diploma dissertation, Humboldt-Universität zu Berlin.