Adapting a part-of-speech tagset to non-standard text: The case of STTS

Heike Zinsmeister, Ulrich Heid, Kathrin Beck
Hamburg University, Hildesheim University, Tübingen University
heike.zinsmeister@uni-hamburg.de, ulrich.heid@uni-hildesheim.de, kbeck@sfs.uni-tuebingen.de

Abstract

The Stuttgart-Tübingen TagSet (STTS) is a de-facto standard for the part-of-speech tagging of German texts. Since its first publication in 1995, STTS has been used in a variety of annotation projects, some of which have adapted the tagset slightly for their specific needs. Recently, the focus of many projects has shifted from the analysis of newspaper text to that of non-standard varieties such as user-generated content, historical texts, and learner language. These text types contain linguistic phenomena that are missing from or are only suboptimally covered by STTS; in a community effort, German NLP researchers have therefore proposed additions to and modifications of the tagset that will handle these phenomena more appropriately. In addition, they have discussed alternative ways of tag assignment in terms of bipartite tags (stem, token) for historical texts and tripartite tags (lexicon, morphology, distribution) for learner texts. In this article, we report on this ongoing activity, addressing methodological issues and discussing selected phenomena and their treatment in the tagset adaptation process.

Keywords: part-of-speech tagset, non-standard text, community effort, German

1. Introduction

1.1. Objectives

The Stuttgart-Tübingen TagSet (STTS: Schiller et al., 1995; Schiller et al., 1999) is a part-of-speech (POS) tagset for German that has gradually become a de-facto standard. It was developed on the basis of newspaper text, but over the past few years, data from other varieties of German have become available that differ in many respects from the "standard written language" that was the original focus: user-generated data (e.g., from online communication media such as Twitter, chats, forums, etc.), texts in dialect, corpora of learner language, and historical corpora. The computational processing of such texts, including POS tagging, is required for opinion mining in user-generated content, the automatic generation of individualized feedback for learners in ICALL applications, and as a basis for the interpretation of historical texts in digital humanities research, among other pursuits.

These types of text contain phenomena that cannot or can only inadequately be classified and annotated using the standard STTS tagset; consequently, interested researchers have come together to create an inventory of critical phenomena and discuss possibilities for additions to or modifications of STTS that will handle these phenomena more appropriately. In this paper, we report the first results (as of Spring 2014) of this initiative.

When standard POS tagging tools based on a standard tagset are applied to texts of the abovementioned types, the results are far from optimal: Certain phenomena are not correctly classified, and the overall tool performanceis comparatively poor (cf. the discussion of POS tagging of web texts in Giesbrecht and Evert, 2009). Although methods for the domain adaptation of trainable tools even for "higher" levels of linguistic analysis (semantic role-labeling, parsing, statistical machine translation) have been frequently discussed (e.g., Daumé III, 2007; Sögaard, 2013), methodological questions concerning the adaptation of tagsets to non-standard genres or types of texts have received much less attention.

In this paper, we will address questions of tagset adaptation to non-standard texts from both a methodological (Section 2) and a practical, applied viewpoint (Section 4). Our examples focus on STTS (Section 3) and proposals for its adaptation to chat data (as an example of computer-mediated communication, CMC), historical texts, and learner texts. In Section 5, we conclude and discuss the next steps planned in the community effort upon which our discussion is based.

1.2. Non-standard texts

The term "non-standard" implies the existence of "standard" texts; in the development of NLP tools, written texts (especially from newspapers) have often been used as raw material for the description and classification of phenomena, as well as for tool training. At the same time, the analysis of news texts was one of the first applications of NLP. Primarily due to the comparative availability of news texts, these implicitly became the "standard". In addition, spoken data (e.g., transcribed interactions) were analyzed and annotation schemes were developed accordingly.

Motivated by newly emerging NLP tasks such as the automatic analysis of user-generated content and intelligent computer-assisted language learning, but also by developments in the digital humanities, NLP research has begun to analyze a wider variety of genres and text types. These include, among others, various types of computer-mediated communication[1], urban youth language[2], but also learner texts[3], and texts from historical stages of the language[4].

Clearly, these inputs vary along many parameters, not only in terms of register, medium, and communicative situation,

[1] http://www.chatkorpus.tu-dortmund.de

[2] http://http://www.kiezdeutschkorpus.de

[3] https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko, http://kobalt-daf.de

[4] http://www.linguistics.ruhr-uni-bochum.de/anselm/

but also with respect to the lexical and grammatical phenomena they contain.

Following the approach of Dipper et al. (2013b), we group this widely divergent material under the term "non-standard text" (or: "non-canonical data") in order to differentiate it from the more commonly investigated newspaper data and from speech data. The NoSta-D corpus[5] is one of the first data collections in the German language that combines different types of non-standard texts, contrasting them with material from newspapers.

In terms of POS annotation, one can either assess the possibilities of using an existing tagset on such non-standard data (cf., e.g., the discussion about different web genres in Giesbrecht and Evert, 2009), or one may explore additions to or modifications of an existing tagset that could better capture the phenomena found in non-standard varieties. As stated above, these varieties are widely divergent in terms of the phenomena that must be addressed: For example, certain genres of computer-mediated communication include emoticons, hashtags, and user addresses ("@john..."), while urban youth language features more spoken-type and code-mixing properties.

Consequently, the objective is not to produce a single POS tagset for all "non-standard" text types, but rather to take stock of the properties of the relevant text types in order to understand what modifications are necessary and how they could be realized. If, following the analysis of several "non-standard" text types, commonalities between some of them can be observed, these can be used in generalizations.

However, at the current stage, we lack clear-cut criteria for text typology and genre delimitation and consequently use sources as an ad-hoc criterion for their classification (e.g. chat, SMS, Twitter, urban youth, learners); in addition, we have not yet finished inventorizing, linguistically describing, and classifying the phenomena found in the data.

But even though the grammar of these different non-standard varieties has yet to be written, we believe that the adaptation of a POS tagset such as STTS to some of them is possible and will help us to better understand the varieties themselves.

2. Issues and methods of tagset design and adaptation

In this section, we review general issues of tagset design as well as more specific aspects concerning the task of tagset adaptation.

2.1. Conceptual and linguistic issues

Part-of-speech tagsets are typically classifications of word forms based on criteria perceived as relevant for the morphosyntactic analysis of texts.

The tags that form part of a tagset are abbreviations of word form types and their features. In actual POS tagging, individual, potentially ambiguous tokens (polysemous or belonging to homonym groups) are considered in context and classified as instances of one of the defined types. A degree of "fuzziness" is inherent in both steps: The type distinctions are not always categorical, instead forming a continuum, and individual tokens may often be interpreted as belonging to one or the other of a pair of (potentially related) types. A prime example is the distinction between adjectival and verbal readings of participle forms (e.g., *begeistert* 'enthusiastic' or the participle of 'inspire').

This issue of "borderline" cases between categories also touches on the granularity of a tagset. If as tagset is very coarse-grained, such that only a few classes are distinguished – in the extreme case, only two classes – it might be possible to avoid ambiguities at the token level. In contrast, if the tagset is very finely-grained (the extreme case would be singleton classes for individual words), distributional peculiarities would not require a word to be classified with different tags; instead, the properties of the specific class would be such that they would model the distribution of the word perfectly, to the exclusion of any other word (for example, in the case of the word "to" in the Penn Treebank tagset; see Santorini, 1990).

Zeldes (to appear) discusses the part-of-speech properties of the German word *voller* 'full of', which exhibits properties of a determiner but can itself co-occur with an article; furthermore, it has peculiar effects on adjective and noun morphology. The question is therefore whether *voller* must be classified as its own singleton part of speech or whether it can be subsumed under the standard schemes of determiners and/or adjectives. Similar questions are discussed by Breindl (2014) with respect to German subordinating and coordinating conjunctions, such as *anstatt* 'instead of', *ausgenommen* 'except', and *es sei denn* 'except'. She shows that it is not possible to apply all defining criteria of conjunctions as specified in Pasch et al. (2003) to these items; in this sense, they are all singletons. The question of tagset granularity is a trade-off between generalizations that group different words together and specializations that model and predict the distribution of individual items.

A related issue concerns the question of which phenomena belong in a POS tagset and which may need to be treated separately: The more clearly the definitional criteria that underlie a given word form type can be determined, the better. A problematic example in this respect is the annotation of "foreign material" (STTS tag "FM") in German texts. The criteria for deciding on the "FM status" (e.g., for *ein cooler*$_{FM/ADJA?}$ *Typ*, 'a cool guy') are far from clear (e.g., etymological vs. functional criteria), and many foreign items have the same function in the sentence as their native counterparts. This may support the treatment of (etymologically) foreign material in a separate layer of annotation rather than within a POS tagset.

2.2. Issues of language engineering

A tagset must be automatically annotatable; with standard statistical tagging, this implies a certain optimal size of the tagset. In addition, the tags must be distinguishable on the basis of properties that are either found in the local context (windows of two to five words) or that can be provided by means of a lexicon – typically distributional or morphological distinctions. It has been shown that semantic dis-

[5] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/clarin-d/standardpage?set_language=en&cl=en

tinctions without morphological or distributional correlates tend to lead to tag confusion errors (Haselbach and Heid, 2010) unless appropriate lexicons or other tools for semantic classification are available; within STTS, this holds, for instance, in the distinction between common nouns (NN) and proper names (NE).

Researchers' work on non-standard texts has raised an additional issue related to the overall architecture of corpus processing. For standard written text, a sequence of tokenizing, POS tagging, and lemmatization followed by parsing is often taken for granted; however, contracted forms (*hamses* = literally, *haben Sie es* 'have you ... it'), erroneous separations of compounds (*Sende Mast* = *Sendemast* 'television tower'), and many other frequent phenomena require either separate normalization steps or a modified repartition of work between tokenizing and POS tagging (cf. Klatt, 2005), as well as raising related questions concerning potential feedback from parsing into tagging (Seeker and Kuhn, 2013).

2.3. Issues of sustainability and interoperability

The above issues apply to both newly designed tagsets and variants of existing ones. More specifically, variants of a tagset may be of different types:

(1) Renaming of categories (isomorphic);

(2) Removal of distinctions (leading to less specificity);

(3) Refinement of distinctions (leading to subtypes of existing types); and

(4) Reclassification (changing the criteria for the delimitation between word form types).

While (1) and (2) are trivial to apply, (3) can be automated only to a certain extent, provided the refinements are identifiable by rules; (4) typically requires manual effort, as existing and new classifications may not be mappable. Changes (1) and (3) can easily be made backward-compatible, but this is less straightforward for (2) and may be very difficult for (4).

Obviously, any tagset adaptation requires the creation of detailed guidelines, including examples and guidance for deciding on easily confusable types. One method of documenting a tagset and supporting semantic interoperability (i.e., mapping between tags from different tagsets at the level of the linguistic distinctions they encode) is to relate the tags in a tagset to data category types from ISOcat[6] and store this mapping in the ISOcat repository. A preliminary mapping from standard STTS tags to ISOcat data categories has been developed, and more detailed documentation on this mapping is being produced. All STTS tags are being integrated into the ISOcat registry, together with their definitions, usage examples, and notes for manual annotation, such as the criteria used to distinguish between two tags.

The STTS interest group discussed whether these tags should be represented as atoms or in a modular way that would mirror their hierarchical structure (see Section 3) and avoid redundancy – for example, independent entries for the class pronoun *P* and its types (definite *D*, indefinite *I*, and their subtypes substitutive *S* and attributive *AT*), with a referencing system to link these to the actual tags (e.g., *PDS*, *PIS*, and *PDAT*). The Dutch tagset for the Corpus Gesproken Nederlands[7] is an example of such a modular representation in ISOcat. However, the STTS group decided against this representation because the non-decomposed form is easier to look up in ISOcat and easier to use. In addition, also from a conceptual point of view, the group rejected a modularized representation. For example, tags such as *NN* and *NE* – normal noun and proper name – have a hierarchical structure, but it is unclear what an ISOcat entry of *E* should be (or the second *N*, for that matter). A slightly different situation is found in the case of the complex tag *APPRART* for contracted prepositions with articles. It is true that it specifies a certain subtype of prepositions (APPR), but the entry of ART cannot be reused to build this subtype, since only definite articles of certain case forms take part in these contractions.

The group plans to list all tagset variants currently in preparation also in ISOcat together with mapping rules to the original set, so that corpora can be more easily compared and mapped.

2.4. Strategies of tagset adaptation: State of the art

The types of procedures listed in Section 2.3 can be used in tagset adaptation. For example, the tagset of the American English Brown Corpus (Francis and Kučera, 1982) was adapted for the Penn Treebank (Santorini, 1990); this process eliminated lexical redundancies and reduced the tagset from 87 categories to only 48 categories. The Penn Treebank tagset itself has recently been adapted for spoken language in the VOICE project through the addition of 26 new spoken discourse-related categories to the original set (VOICE Project, 2013). Other tagset adaptations for English include different versions of the BNC's CLAWS tagset (Garside, 1996), the application of the SUSANNE tagset to the (spoken) CHRISTINE corpus (Rahman and Sampson, 1999), and the application of the TOSCA tagset to the ICE corpora (Aarts, 1992).

A typical case of (rather radical) reduction in granularity is the Universal Tagset (Petrov et al., 2012), which seeks to provide a language-independent set of only twelve universal tags. The motivation for this reduction is to create a solid basis for cross-linguistic analysis that can be used for multilingual parsing and automatic translation. The tagset is reduced to only major POS categories and does not distinguish between distributional differences; for example, it does not distinguish between subordinating and coordinating conjunctions.

For German, a number of variants of STTS have been proposed for the annotation of written texts, often in response to tag confusion problems; examples include the guidelines for the Tiger corpus (Albert et al., 2003: changes of types 1, 2, and 3, cf. Section 2.3.) and those of TüBa-D/Z (Telljohann et al. 2012: type 1), or of the Zürich-based system UIS (Schneider, online: types 2 and 3). A methodologically

[6] http://www.isocat.org

[7] http://lands.let.ru.nl/cgn/home.htm

different approach has been applied in the design of the tagset HiTS, which is intended for the annotation of historical texts (Dipper et al., 2013a): The tagset was designed on the basis of the phenomena of historical stages of German, but retains the distinctions of STTS. As a result, mappings from HiTS to STTS have been proposed wherever possible; this basic compatibility allows users to jointly query texts from different time periods that are annotated partly in HiTS and partly in STTS, but that exist together in one query system.

A completely different approach to tagset adaptation is discussed in Díaz-Negrillo et al. (2010). Instead of reducing or extending the tagset, the authors suggest a modularization of the tags into their three defining dimensions of lexicon, morphology, and distribution (cf. Section 4.5.2 for a discussion).

3. Characteristics of STTS

The standard version of STTS has 54 tags. STTS is organized as a logical tagset (cf., Leech 1997, p. 33); its categories (e.g., ADJ for adjectives) have subtypes (e.g., ADJA for attributive and ADJD for predicative adjectives). The tags can be mapped onto attribute-value pairs (as recommended by EAGLES),[8] and further morphosyntactic attributes can be added as a second layer (cf. RFTagger, Schmid and Laws, 2008) that specializes the annotation on the basis of lexical information (*NN* vs. $NN_{sing.dat.mask}$), for example.

In principle, this property of STTS permits variable-depth additions to the tagset: In a given STTS subtype, one category may be further subdivided, while this same category can be used elsewhere with its major class only. Variable-depth annotation has been applied, for example, in the annotation of discourse relation senses (Prasad et al., 2007) and preposition senses (Müller et al., 2010).

The main word classes distinguished by STTS are summarized in Table 1.

1.	Nouns (N)	7.	Adverbs (ADV)
2.	Verbs (V)	8.	Conjunctions (KO)
3.	Articles (ART)	9.	Adpositions (AP)
4.	Adjectives (ADJ)	10.	Interjections (ITJ)
5.	Pronouns (P)	11.	Particles (PTK)
6.	Cardinals (CARD)		

Table 1: Major classes of STTS (Schiller et al., 1999, p. 4)

The STTS tag types are defined according to the lexical, morphological, and distributional properties of items, and more than one of these criteria may apply to a given distinction. The distinction between full verbs (VV...) and modal or temporal auxiliaries (VM..., VA...) is lexical, as the latter two classes can be enumerated in full in the tagger lexicon; the distinction between attributive and predicative adjectives, *das große*$_{ADJA}$ *Haus*, 'the big house' vs. *das Haus ist groß*$_{ADJD}$ 'the house is big', is morphologically marked (ADJA being inflected, whereas ADJD is not) and at the same time characterized by different distributions (prenominal vs. close to a copula).

[8] http://www.ilc.cnr.it/EAGLES/annotate/annotate.html

4. Examples of tagset adaptation for non-standard text

4.1. Workflow

In the interest group on STTS adaptation, the following overall workflow has been designed for the definition of additions to and modifications of the existing tagset:

1. Identification of relevant language phenomena that must be captured, e.g., on the basis of error analyses;
2. Subclassification of the phenomena from (1) in terms of properties that may be used as tag distinction criteria; proposal of appropriate tags;
3. Tests of annotation accuracy, both manual (via measurements of inter-annotator agreement) and automatic (via intrinsic or extrinsic evaluation, i.e., in terms of measurements of tagging accuracy against a gold standard or by means of a task-based evaluation assessing the quality of applications that depend on POS input such as parsing);
4. Possibly iterative improvements, consensus-building, and ISOcat documentation.

As of March 2014, steps (1) and (2) have been completed for chat texts (taken as a subtype of computer-mediated communication, CMC), and step (3) has been realized in part for learner data and for urban youth language. The completion of step (3) and step (4) for CMC data is planned for 2015.

4.2. Sample phenomena from non-standard texts

The discussion below of the phenomena to be addressed in the POS tagging of certain types of non-standard texts is by no means exhaustive; these examples are cited because we use them in Section 4.3 to illustrate our procedures in tagset adaptation.

4.2.1. Phenomena from computer-mediated communication

The following discussion is based on analyses of the Dortmund chat corpus[9] (Bartz et al., 2013b) and on proposals presented in Bartz et al. (2013a).

The chat data include specific symbols: emoticons (represented as ASCII symbol sequences (e.g., ":-)") or as graphical icons) and hashtags or addresses in connecting with items (*Kreta war super! #Urlaub:* 'Crete was great! #holidays'; *@lothar: wie isset so?* '@lothar: how is it going?'). When the latter two are used in a syntactically non-integrated manner, as in the two examples above, they are interpreted as having their own (communicative) meaning and are consequently assigned their own tags. Similarly, URLs are annotated separately.

The above phenomena do not or only very rarely occur in standard texts; moreover onomatopoetic items (*boing, miau, tok*) are more frequent in CMC data than in other text

[9] http://www.chatkorpus.tu-dortmund.de

types. These items are classified as interjections in standard STTS, but due to their distributional behavior in chat data, it may more appropriate to consider them as a separate category.

Another CMC-specific item involves "action words" (Aktionswörter), which in German often appear as uninflected verb stems (**lach** 'laughing', *freu* 'happy', **lol**). These may contain elements other than verbs (which is why it would be inappropriate to classify them as verb forms, cf. **lol**), and they tend to be syntactically non-integrated, often appearing as comments on previous text.

Finally, CMC data also contain many contraction forms, see Section 2.3 for preposition + article forms: *zum* (= *zu dem*), *ins* (= *in das*), etc. Like spoken language, chat very frequently uses contractions of verb forms and personal pronouns: *schreibste, machste* for *schreibst du*, *machst du* ('you write', 'you make').

In a sample of 118 698 tokens from the Dortmund chat corpus, 664 contractions forms were counted, of which verb+pronoun cases made up for about 64%, and contractions with articles as second element for about 30%.

In CMC data we also find significant quantities of discourse markers. Experiments have shown that their classification in STTS was suboptimal for the annotation of CMC data, as their distribution over the classes of adverbs, particles, and interjections was not based on criteria that can be verified in surface syntax.

4.2.2. Phenomena from learner language

Reznicek and Zinsmeister (2013) discuss phenomena in texts from learners of German as a foreign language, in which tagging with STTS resulted in conflicting outcomes. This was due to the fact that learners deviate in their language use from the target lexicon, morphology, and/or distribution, such that different dimensions of tag definitions point to different tags. In (1), the morphology required a participle, whereas the distribution pointed to an infinitive – and the lexical dimension did not favor either one.

(1) Wenn bei mir etwas passierte, kann ich dass mit meinen Eltern besprochen[→ besprechen].
'If somethings happens to me, I can talked[→ talk] about it with my parents.'

4.2.3. Phenomena from historical texts

Dipper et al. (2013a) discuss in detail phenomena from historical texts that have required adaptation of the POS tagset. We will only describe some of these cases here. The Modern German negation marker *nicht* 'not' emerged from an indefinite pronoun that co-occurred with an independent negation particle. This type of pronoun no longer exists in Modern German, so there is no part-of-speech tag available in STTS that captures its distribution. Furthermore, adjectives used to have more flexible distribution in earlier stages of German than they do today. In addition to prenominal modification, postnominal modification was common, although it is found only very rarely in Modern German (*Hänschen klein* 'John little'). In addition, on a regular basis adjectives were substituted for nominals in cases that would either be ellipsis or (capitalized) nominalizations in Modern German (*die rehten skinent* vs. *die gerechten/Gerechten strahlen* 'the righteous$_{N/ADJ}$ are beaming'). Finally, punctuation has changed over the course of time. Old High German used periods almost exclusively, but not necessarily in their modern function marking the ends of sentences. Other punctuation signs and markers mostly emerged during the Middle High German period, but again, they were not necessarily used in the same way as they are today. Hence, the modern distinction between sentence final punctuation (typically periods, question marks, exclamation marks) and sentence-internal markers (typically commas) does not make sense in the context of historical documents.

4.3. STTS adaptation proposals

4.3.1. Proposals for CMC data

The data discussed, in Section 4.2.1, require tagset modifications of types (3) and (4) (Section 2.3) to be handled appropriately.

Simple additions based on formal grounds (which may interact with tokenizing to some extent) concern tags for emoticons (EMO), hashtags (HST), addressing markers (ADR), and URLs (URL). Bartz et al. (2013a) also encourage the introduction of tags for contraction forms. We are still discussing whether a single generic tag will be sufficient or whether individual tags for each type of contracted form should be introduced. Alternatively, a separate step of tokenizing (and/or normalizing) might be preferable, e.g., for the training of statistical tools in order to avoid data sparseness; this would separate, for example, *haste* into *hast du* 'you have' while retaining (in a different annotation layer) the information that the items were combined into one string.

Onomatopoetic items have been reclassified from interjections (ITJ) to a separate word class (ONO). Bartz et al. (2013b) also suggest introducing a tag (AW) for action words, thus far classified in an arbitrary way by automatic taggers, e.g., by TreeTagger (Schmid, 1995).

As stated in Section 4.2.1, the annotation of CMC data requires the adequate treatment of discourse markers. On a related note, a major reclassification of adverbs, particles, and interjections has been proposed. In fact, standard STTS differentiates between adverbs (ADV: non-inflectable, e.g., *vermutlich* 'probably'), grammatical particles (PTK.*: mainly closed sets, but also including, e.g., *ja* 'yes' and *nein* 'no') and interjections. Based on Hirschmann et al. (2013), a simple distinction was proposed to separate items that can be heads of phrases with grammatical functions (adverbs: ADV) from items that cannot (particles: PTK.*).

Bartz et al. (2013a) have furthermore suggested that syntactically unintegrated discourse markers be analyzed as a separate class ("selbstständige Interaktive Einheiten" 'autonomous interaction elements', SIE).

The SIE category has functional subtypes (interjections, response items, (non-inflecting) action words, pause fillers, onomatopoetic items, emoticons), some of which in turn have distributionally defined subtypes according to their position in a sentence or a turn (initial, internal, final, isolated, cf. Bartz et al., 2013a). Grouping all interaction items under SIE would permit more coherent and better repro-

ducible annotation also of adverbs (temporal, local, and modal) and particles (only grammatical ones).

4.3.2. Proposals for learner language

Díaz-Negrillo et al. (2010) have introduced a compelling proposal for the POS tagging of learner language. They suggest the modularization of tagging into the three dimensions of lexicon, morphology, and distribution; when the different dimensions suggest divergent interpretations, tags may be assigned in feature bundles (e.g., VVINF-VVPP-VVINF would indicate that the lexicon suggests VVINF, morphology VVPP, and distribution VVINF). This development is based on the assumption that these three dimensions are relevant for a learner's interlingua, which might diverge from the target structure along these lines. When a dimension does not single out one category, ambiguous tags are assigned.

Reznicek and Zinsmeister (2013) proposed a different type of multi-dimensionality for the POS tagging of learner texts. They suggested merging POS tags from the learner text with POS tags of an aligned normalization level, in order to capture divergences in the learner language without the need to be explicit with regard to the predictions of the three dimensions (see above). Their portemanteau tags have a procedural nature, as they require the tagging of both text levels before they can be merged. To capture common mismatches, the authors also discuss the option of using more general meta-classes; however, the modeling of learners' mismatch classes does not necessarily conform with the abstraction hierarchy implicitly encoded in STTS. Further task-based experiments will be necessary to determine whether the generalization across mismatch classes is useful.

4.3.3. Proposals for historical texts

The HiTS tagset (Dipper et al., 2013a) models phenomena of historical language that cannot be captured by simply applying STTS. First, it includes additional tags for types that do not longer exist in Modern German or only occur very rarely (e.g., PNEG for negative-polar pronouns, ADJN for postnominal attributive adjectives; see also Section 4.2.3). These types can be mapped to modern "versions" (e.g., indefinite pronouns PIS and attributive adjectives ADJA) with a certain degree of information loss. As well as adding new subtypes, HiTS also employs a kind of multi-dimensional approach, in that it distinguishes between the POS of the lexical stem and the POS of the token (in running text). This results in a bipartite POS tag that often captures diachronic development (e.g., *niouuiht*/PI >PNEG encodes that the word used to be an indefinite pronoun and is now used as a negative (indefinite) pronoun).

Some HiTS tags correspond to more than one STTS tag: This means that they cannot simply be mapped, but instead require rules to decide which STTS tag should be chosen. A simple case is punctuation, i.e., the tokens . *:* , *?* *!*, which have changed function over the course of time. In HiTS, these tokens are marked with the general punctuation tag $_; in STTS they are divided into sentence-final and sentence-internal punctuation based on their form ($. vs. $,). In other cases, it is not possible to define a disambiguation rule. For example, there is no direct way to decide whether an instance of substituting attributive adjective (ADJS; see Section 4.2.3) corresponds to Modern German noun ellipsis (ADJA) or nominalization (NN). In such cases, the mapping guidelines must provide a general rule of thumb for how the tag should be mapped.

An additional aspect of HiTS is that it must reflect traditions in historical linguistics that have been accepted by the community. For instance, the introduction of the tag NA ("appellative noun"), which can be isomorphically mapped to NN ("normal noun"), was motivated by such considerations.

4.4. Evaluation of POS tagging of non-standard texts

Standard methods of evaluating the adequacy and reliability of POS tagsets include (i) manual annotation and inter-annotator agreement measures, (ii) automatic annotation and accuracy evaluation against a gold standard, and (iii) task-based evaluation, e.g., in the context of parsing.

The proposals discussed in this paper are too recent for such evaluations to be available. However, there have been assessments of the performance of standard taggers trained on news text using standard STTS on certain non-standard text types.

State-of-the-art tagging results on newspaper text (Tiger corpus, Brants et al., 2004) with STTS are in the range of 96% to 98% accuracy, depending on the tagger used. When the tagger is applied to web data (DeWaC corpus, Baroni and Kilgarriff, 2006) the accuracy rate drops to 90.87% to 93.71% without tagset adaptation (Giesbrecht and Evert, 2009).

On orthographically transcribed data from the spoken FOLK corpus, Westpfahl and Schmidt (2013) report a per-token accuracy of 81.16% using standard TreeTagger and standard STTS. The largest proportion of tagging errors involved (target) particles and interjections, followed by pronouns and verbs. The researchers conclude that there is a need to improve the modeling of particles in STTS (see Section 4.3.1).

For learner texts, Reznicek and Zinsmeister (2013) argue that it is not possible to create a gold standard for POS tagging based on the learner text itself, at least not for a one-dimensional tagset like STTS (cf. Section 4.3.2). They follow Reznicek et al. (2013) in creating parallel normalization layers that are aligned to the learner text at the token level. These normalization layers are used in the development of a POS gold standard against which the learner tags can be evaluated. In a small study, the researchers evaluated an ensemble tagger consisting of three (off-the-shelf) taggers using TreeTagger (Schmid, 1995) as default. The per-token accuracy for the L2 texts ranged between 94.8% and 96.9%, whereas accuracy on the normalized texts was between 97.0% and 97.8%. The effect can only partially be attributed to the L2 status, a control group of German L1 texts was tagged with only 95.6% accuracy (and 96.5% for the normalized form). However, sentence length had an effect; in addition, the control texts should be classified as non-standard language, since they were argumentative essays written by high-school students.

The only case of an evaluation on non-standard text based

on an adapted German tagset that we are aware of is Rehbein and Schalowski's (2013) work on transcribed spoken urban youth language. Using standard STTS and TreeTagger (Schmid, 1995) trained on news text, they report a baseline of 42.48% per-token accuracy.

To improve performance, they extended STTS with eleven tags related to spoken language (pause, filler, backchannel signal, some particle subtypes, a tag for uninterpretable material, and one for unfinished utterances). Annotation experiments with human annotators determined that the extended tagset can be assigned with high reliability. In a preliminary experiment with three human annotators, they obtained 96.5% per-token accuracy and Fleiss' κ of 0.075 on a small test set, which is close to the results achieved for annotating written text with standard STTS: Rehbein et al. (2012) report 97.9% per-token accuracy and κ of 0.979 for two human annotators on normalized (i.e., grammatical correct) learner texts. In a more in-depth annotation experiment incorporating the newly introduced (discourse) particle subtypes, two annotators achieved 88.20% agreement on average (ranging from 45.45% to 96.87%, including the standard STTS answer particle PTKANT at 89.59%).

Re-training of TreeTagger on in-domain data with the extended tagset increased the tagger's accuracy from 42.48% (see above) to 59.90%. Rehbein et al. also developed their own tagger based on Conditional Random Fields, which achieved an accuracy rate of up to 91.09% when additional domain adaptation techniques were applied to reduce the amount of out-of-vocabulary data.

5. Conclusion – Further Work

In this article, we have presented the Stuttgart-Tübingen tagset (STTS, a quasi-standard POS tagset for German-language texts), addressing issues that arise in the adaptation of this tagset to non-standard text data, such as user-generated texts from online communication media, learner corpora, and historical corpora.

We have reported on community efforts in adaptation; the next steps to be undertaken will also be based on broad consensus-building activities. For some types of non-standard texts, a detailed analysis of new or "non-canonical" phenomena has been offered. More such work is needed: For example, Bartz et al. (2013a) have offered elements for the analysis of transcribed spoken data. On the basis of such work, proposals for adapted versions of the STTS tagset have been made. As of the first half of 2014, data are being prepared for shared tasks on the annotation of samples from the Dortmund chat corpus. Manual annotation (and the identification of inter-annotator agreement) in combination with automatic annotation and accuracy evaluations will determine which of the proposed changes merit further development in order to achieve high-quality annotation of these types of texts.

In the medium term, it will also be interesting to assess whether the tagset additions introduced for CMC data can be used for transcribed speech as well. In this way, we intend to arrive at generalizations that not only provide coherent tagset adaptations for individual types of non-standard texts, but also allow us to compare different non-standard text types.

To ensure sustainability and reproducibility, we will finalize the mapping of standard STTS to ISOcat data categories before mapping the existing and upcoming variants in a similar manner, so that we will be able to relate standard STTS and its variants and extensions through the RELcat mechanism (Windhouwer, 2012). In addition, we will link the classifications underlying STTS and its variants to categories and terminology from descriptive linguistics in order to facilitate the training of non-computational linguists (e.g., in the digital humanities).

Finally, the community efforts in adapting STTS to non-standard data are documented on a wiki page, which provides an overview of STTS, its variants, and the ongoing work for all interested readers.[10]

Acknowledgments

This paper summarizes the discussions from three workshops on STTS organized by the authors of this paper in 2012 and 2013. We would like to thank Swantje Westpfahl and Thomas Schmidt (Mannheim), Thomas Bartz, Michael Beißwenger, and Angelika Storrer (Dortmund), Ines Rehbein (Potsdam), and Stefanie Dipper and Julia Krasselt (Bochum) for their detailed input to Section 4; moreover, we thank all contributors to the three workshops for their many inspiring discussions and comments.

6. References

Aarts, J. (1992). Comments to "A new corpus of English: ICE" by Sidney Greenbaum. In Svartvik, J., editor, *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, pages 180–183. Walter de Gruyter, Berlin.

Albert, S., Anderssen, J., Bader, R., Becker, S., Bracht, T., Brants, S., Brants, T., Demberg, V., Dipper, S., Eisenberg, P., Hansen, S., Hirschmann, H., Janitzek, J., Kirstein, C., Langner, R., Michelbacher, L., Plaehn, O., Preis, C., Pußel, M., Rower, M., Schrader, B., Schwartz, A., Smith, G., and Uszkoreit, H., (2003). *TIGER Annotationsschema*. Saarland University, University of Stuttgart, and University of Potsdam.

Bartz, T., Beißwenger, M., Rehbein, I., Schmidt, T., Storrer, A., and Westpfahl, S. (2013a). Modifikation und Erweiterung von STTS für die Annotation von Gesprächskorpora und von Korpora zu Genres internetbasierter Kommunikation. Presentation at GSCL Kaleidoskop-2013, Darmstadt, Germany.

Bartz, T., Beißwenger, M., and Storrer, A. (2013b). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, 28(1):155–198.

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

[10] STTS wiki: `http://www.linguistics.ruhr-uni-bochum.de/stts`; all URLs in this paper have been checked as of 03/21/2014.

Breindl, E. (2014). Gemeinsam einsam: Was eint satzverknüpfende Einzelgänger? Presentation at DGfS-2014, Marburg, Germany.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 256–263, Prague, Czech Republic.

Díaz-Negrillo, A., Meurers, W., Valera, S., and Wunsch, H. (2010). Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154.

Dipper, S., Donhauser, K., Klein, T., Linde, S., Müller, S., and Wegera, K.-P. (2013a). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *JLCL*, 28(1):85–137.

Dipper, S., Lüdeling, A., and Reznicek, M. (2013b). NoSta-D: A corpus of German non-standard varieties. In Zampieri, M. and Diwersy, S., editors, *Non-standard Data Sources in Corpus-based Research*, pages 69–76. Shaker Verlag, Aachen.

Francis, W. and Kučera, H. (1982). *Frequency Analysis of English Usage. Lexicon and Grammars*. Houghton Mifflin, Boston.

Garside, R. (1996). The robust tagging of unrestricted text: The BNC experience. In Thomas, J. and Short, M., editors, *Using corpora for language research: Studies in the Honour of Geoffrey Leech*, pages 167–180. Longman, London.

Giesbrecht, E. and Evert, S. (2009). Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In *Proceedings of WAC5*, San Sebastian, Spain.

Haselbach, B. and Heid, U. (2010). The Development of a Morphosyntactic Tagset for Afrikaans and its Use with Statistical Tagging. In *Proceedings of LREC-2010*, Valletta, Malta.

Hirschmann, H., Lestmann, N., Rehbein, I., and Westpfahl, S. (2013). Erweiterung der Wortartenkategorien des STTS im Bereich 'ADV' und 'PTK…'. Presentation at STTS Workshop, Hildesheim, Germany.

Klatt, S. (2005). *Kombinierbare Textanalyseverfahren für die Korpusannotation und Informationsextraktion*. Shaker Verlag, Aachen.

Müller, A., Hülscher, O., Roch, C., Keßelmeier, K., Stadtfeld, T., Strunk, J., and Kiss, T. (2010). An annotation schema for preposition senses in German. In *Proceedings of LAW IV*, pages 177–181, Uppsala, Sweden.

Pasch, R., Braußе, U., Breindl, E., and Waßner, U. (2003). *Handbuch der deutschen Konnektoren*. De Gruyter, Berlin, New York.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of LREC-2012*, pages 2089–2096, Istanbul, Turkey.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). The Penn Discourse Treebank 2.0 annotation manual.

Rahman, A. and Sampson, G. (1999). Extending grammar annotation standards to spontaneous speech. In Kirk, J., editor, *Corpora Galore: Analyses and Techniques in Describing English*. Rodopi, Amsterdam.

Rehbein, I. and Schalowski, S. (2013). STTS goes Kiez – Experiments on Annotating and Tagging Urban Youth Language. *JLCL*, 28(1):199–227.

Rehbein, I., Hirschmann, H., Lüdeling, A., and Reznicek, M. (2012). Better tags give better trees or do they? *LiLT*, 7(10).

Reznicek, M. and Zinsmeister, H. (2013). STTS-Konfusionsklassen beim Tagging von Fremdsprachlernertexten. *JLCL*, 28(1):63–83.

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In Ballier, N., Díaz Negrillo, A., and Thompson, P., editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123. John Benjamins, Amsterdam.

Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Technical Report 3rd Revision, 2nd Printing, University of Pennsylvania.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textcorpora. Technical report, University of Stuttgart and University of Tübingen.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of Coling 2008)*, pages 777–784, Manchester, UK.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Schneider, G. (online). STTS-Tagset (Züricher Variante). `https://files.ifi.uzh.ch/CL/tagger/UIS-STTS-Diffs.html`.

Seeker, W. and Kuhn, J. (2013). Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39:23–55.

Sögaard, A. (2013). *Semi-Supervised Learning and Domain Adaptationin Natural Language Processing*. Synthesis Lectures on Human Language Technology. Morgan & Claypool.

Telljohann, H., Hinrichs, E., Kübler, S., Zinsmeister, H., and Beck, K. (2012). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, University of Tübingen.

VOICE Project. (2013). VOICE Part-of-Speech Tagging and Lemmatization Manual. Technical report, Universiy of Vienna.

Westpfahl, S. and Schmidt, T. (2013). POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *JLCL*, 28(1):139–154.

Windhouwer, M. (2012). RELcat: a Relation Registry for ISOcat data categories. In *Proceedings of LREC-2012*, pages 3661–3664, Istanbul, Turkey.

Zeldes, A. (to appear). The case for caseless prepositional constructions with *voller* in German. In Boas, H. C. and Ziem, A., editors, *Constructional Approaches to Argument Structure in German*. De Gruyter, Berlin.