Towards robust multi-tool tagging. An OWL/DL-based approach

Christian Chiarcos
University of Potsdam, Germany
chiarcos@uni-potsdam.de

Abstract

This paper describes a series of experiments to test the hypothesis that the parallel application of multiple NLP tools and the integration of their results improves the correctness and robustness of the resulting analysis.

It is shown how annotations created by seven NLP tools are mapped onto tool-independent descriptions that are defined with reference to an ontology of linguistic annotations, and how a majority vote and ontological consistency constraints can be used to integrate multiple alternative analyses of the same token in a consistent way.

For morphosyntactic (parts of speech) and morphological annotations of three German corpora, the resulting merged sets of ontological descriptions are evaluated in comparison to (ontological representation of) existing reference annotations.

1 Motivation and overview

NLP systems for higher-level operations or complex annotations often integrate redundant modules that provide alternative analyses for the same linguistic phenomenon in order to benefit from their respective strengths and to compensate for their respective weaknesses, e.g., in parsing (Crysmann et al., 2002), or in machine translation (Carl et al., 2000). The current trend to parallel and distributed NLP architectures (Aschenbrenner et al., 2006; Gietz et al., 2006; Egner et al., 2007; Luís and de Matos, 2009) opens the possibility of exploring the potential of redundant parallel annotations also for lower levels of linguistic analysis.

This paper evaluates the potential benefits of such an approach with respect to morphosyntax (parts of speech, pos) and morphology in German: In comparison to English, German shows a rich and polysemous morphology, and a considerable number of NLP tools are available, making it a promising candidate for such an experiment.

Previous research indicates that the integration of multiple part of speech taggers leads to more accurate analyses. So far, however, this line of research focused on tools that were trained on the same corpus (Brill and Wu, 1998; Halteren et al., 2001), or that specialize to different subsets of the same tagset (Zavrel and Daelemans, 2000; Tufiş, 2000; Borin, 2000). An even more substantial increase in accuracy and detail can be expected if tools are combined that make use of different annotation schemes.

For this task, ontologies of linguistic annotations are employed to assess the linguistic information conveyed in a particular annotation and to integrate the resulting ontological descriptions in a consistent and tool-independent way. The merged set of ontological descriptions is then evaluated with reference to morphosyntactic and morphological annotations of three corpora of German newspaper articles, the NEGRA corpus (Skut et al., 1998), the TIGER corpus (Brants et al., 2002) and the Potsdam Commentary Corpus (Stede, 2004, PCC).

2 Ontologies and annotations

Various repositories of linguistic annotation terminology have been developed in the last decades, ranging from early texts on annotation standards (Bakker et al., 1993; Leech and Wilson, 1996) over relational data base models (Bickel and Nichols, 2000; Bickel and Nichols, 2002) to more recent formalizations in OWL/RDF (or with OWL/RDF export), e.g., the General Ontology of Linguistic Description (Farrar and Langendoen, 2003, GOLD), the ISO TC37/SC4 Data Category Registry (Ide and Romary, 2004; Kemps-

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 659–670,
Uppsala, Sweden, 11-16 July 2010.

Snijders et al., 2009, DCR), the OntoTag ontology (Aguado de Cea et al., 2002), or the Typological Database System ontology (Saulwick et al., 2005, TDS). Despite their common level of representation, however, these efforts have not yet converged into a unified and generally accepted ontology of linguistic annotation terminology, but rather, different resources are maintained by different communities, so that a considerable amount of disagreement between them and their respective definitions can be observed.[1]

Such conceptual mismatches and incompatibilities between existing terminological repositories have been the motivation to develop the OLiA architecture (Chiarcos, 2008) that employs a shallow Reference Model to mediate between (ontological models of) annotation schemes and several existing terminology repositories, incl. GOLD, the DCR, and OntoTag. When an annotation receives a representation in the OLiA Reference Model, it is thus also interpretable with respect to other linguistic ontologies. Therefore, the findings for the OLiA Reference Model in the experiments described below entail similar results for an application of GOLD or the DCR to the same task.

2.1 The OLiA ontologies

The Ontologies of Linguistic Annotations – briefly, OLiA ontologies (Chiarcos, 2008) – represent an architecture of modular OWL/DL ontologies that formalize several intermediate steps of the mapping between concrete annotations, a Reference Model and existing terminology repositories ('External Reference Models' in OLiA terminology) such as the DCR.[2]

The OLiA ontologies were originally developed as part of an infrastructure for the sustainable maintenance of linguistic resources (Schmidt et al., 2006) where they were originally applied to the formal representation and documentation of annotation schemes, and for concept-based annotation queries over to multiple, heterogeneous corpora annotated with different annotation schemes (Rehm et al., 2007; Chiarcos et al., 2008). NLP applications of the OLiA ontologies include a proposal to integrate them with the OntoTag ontologies and to use them for interface specifications between modules in NLP pipeline architectures (Buyko et al., 2008). Further, Hellmann (2010) described the application of the OLiA ontologies within NLP2RDF, an OWL-based blackboard approach to assess the meaning of text from grammatical analyses and subsequent enrichment with ontological knowledge sources.

OLiA distinguishes three different classes of ontologies:

- The OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is primarily based on a blend of concepts of EAGLES and GOLD, and further extended in accordance with different annotation schemes, with the TDS ontology and with the DCR (Chiarcos, 2010).
- Multiple OLiA ANNOTATION MODELs formalize annotation schemes and tag sets. Annotation Models are based on the original documentation and data samples, so that they provide an authentic representation of the annotation not biased with respect to any particular interpretation.
- For every Annotation Model, a LINKING MODEL defines `subClassOf` ($\sqsubseteq$) relationships between concepts/properties in the respective Annotation Model and the Reference Model. Linking Models are interpretations of Annotation Model concepts and properties in terms of the Reference Model, and thus multiple alternative Linking Models for the same Annotation Model are possible. Other Linking Models specify $\sqsubseteq$ relationships between Reference Model concepts/properties and concepts/properties of an External Reference Model such as GOLD or the DCR.

The OLiA Reference Model (namespace `olia`) specifies concepts that describe linguistic categories (e.g., `olia:Determiner`) and grammatical features (e.g., `olia:Accusative`), as well

[1] As one example, a GOLD Numeral is a Determiner (Numeral $\sqsubseteq$ Quantifier $\sqsubseteq$ Determiner, `http://linguistics-ontology.org/gold/2008/Numeral`), whereas a DCR Numeral is defined on the basis of its semantic function, without any references to syntactic categories (`http://www.isocat.org/datcat/DC-1334`). Thus, *two* in *two of them* is a DCR Numeral but not a GOLD Numeral.

[2] The OLiA Reference Model is accessible via `http://nachhalt.sfb632.uni-potsdam.de/owl/olia.owl`. Several annotation models, e.g., `stts.owl`, `tiger.owl`, `connexor.owl`, `morphisto.owl` can be found in the same directory together with the corresponding linking files `stts-link.rdf`, `tiger-link.rdf`, `connexor-link.rdf` and `morphisto-link.rdf`.

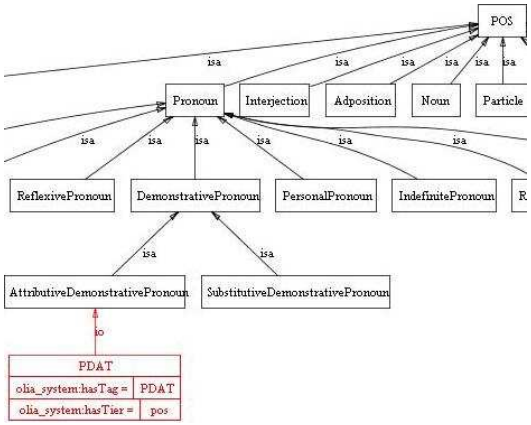

Figure 1: Attributive demonstrative pronouns (`PDAT`) in the STTS Annotation Model

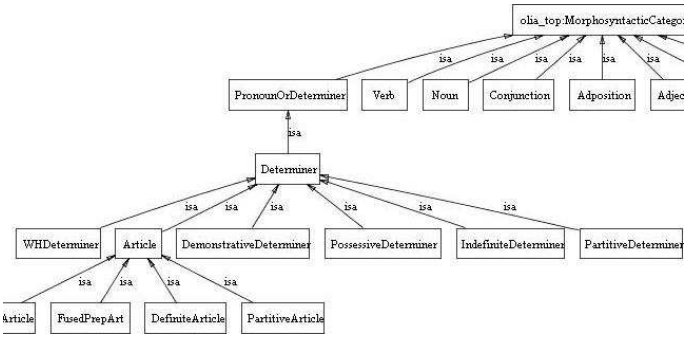

Figure 2: Selected morphosyntactic categories in the OLiA Reference Model

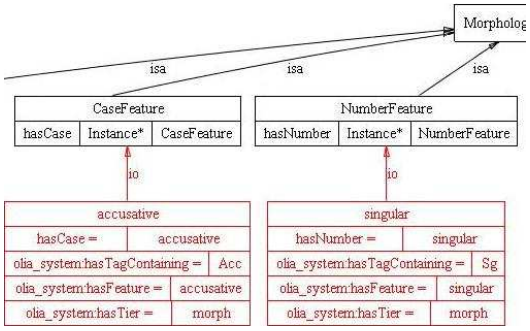

Figure 3: Individuals for accusative and singular in the TIGER Annotation Model

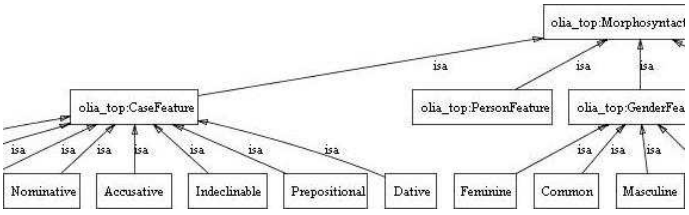

Figure 4: Selected morphological features in the OLiA Reference Model

as properties that define possible relations between those (e.g., `olia:hasCase`). More general concepts that represent organizational information rather than possible annotations (e.g., `MorphosyntacticCategory` and `CaseFeature`) are stored in a separate ontology (namespace `olia_top`).

The Reference Model is a shallow ontology: It does not specify disjointness conditions of concepts and cardinality or domain restrictions of properties. Instead, it assumes that such constraints are inherited by means of $\sqsubseteq$ relationships from an External Reference Model. Different External Reference Models may take different positions on the issue – as languages do[3] –, so that this aspect is left underspecified in the Reference Model.

Figs. 2 and 4 show excerpts of category and feature hierarchies in the Reference Model.

With respect to morphosyntactic annotations (parts of speech, `pos`) and morphological annotations (`morph`), five Annotation Models for German are currently available: STTS (Schiller et al., 1999, `pos`), TIGER (Brants and Hansen, 2002, `morph`), Morphisto (Zielinski and Simon, 2008, `pos, morph`), RFTagger (Schmid and Laws, 2008, `pos, morph`), Connexor (Tapanainen and Järvinen, 1997, `pos, morph`). Further Annotation Models for `pos` and `morph` cover five different annotation schemes for English (Marcus et al., 1994; Sampson, 1995; Mandel, 2006; Kim et al., 2003, Connexor), two annotation schemes for Russian (Meyer, 2003; Sharoff et al., 2008), an annotation scheme designed for typological research and currently applied to approx. 30 different languages (Dipper et al., 2007), an annotation scheme for Old High German (Petrova et al., 2009), and an annotation scheme for Tibetan (Wagner and Zeisler, 2004).

[3]Based on primary experience with Western European languages, for example, one might assume that a `hasGender` property applies to nouns, adjectives, pronouns and determiners only. Yet, this is language-specific restriction: Russian finite verbs, for example, show gender congruency in past tense.

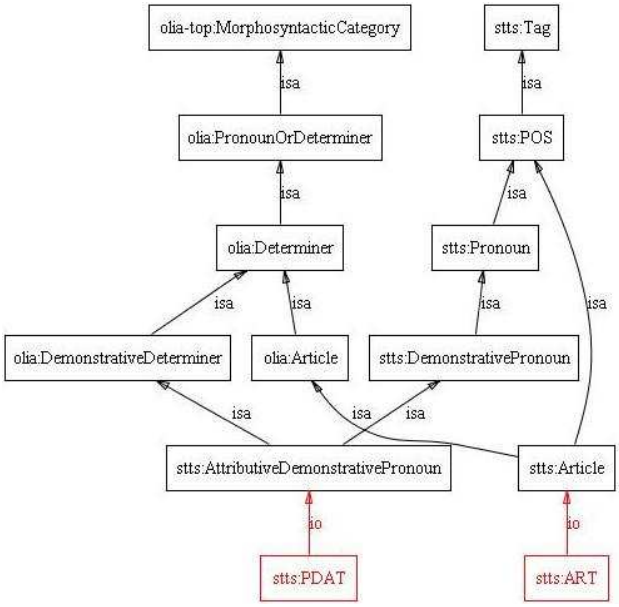

Figure 5: The STTS tags `PDAT` and `ART`, their representation in the Annotation Model and linking with the Reference Model.

Annotation Models differ from the Reference Model mostly in that they include not only concepts and properties, but also individuals: Annotation Model concepts reflect an abstract conceptual categorization, whereas individuals represent concrete values used to annotate the corresponding phenomenon. An individual is applicable to all annotations that match the string value specified by this individual's `hasTag`, `hasTagContaining`, `hasTagStartingWith`, or `hasTagEndingWith` properties. Fig. 1 illustrates the structure of the STTS Annotation Model (namespace `stts`) for the individual `stts:PDAT` that represents the tag used for attributive demonstrative pronouns (demonstrative determiners). Fig. 3 illustrates the individuals `tiger:accusative` and `tiger:singular` from the hierarchy of morphological features in the TIGER Annotation Model (namespace `tiger`).

Fig. 5 illustrates the linking between the STTS Annotation Model and the OLiA Reference Model for the individuals `stts:PDAT` and `stts:ART`.

2.2 Integrating different morphosyntactic and morphological analyses

With the OLiA ontologies as described above, annotations from different annotation schemes can now be interpreted in terms of the OLiA Reference Model (or External Reference Models like GOLD or the DCR).

As an example, consider the attributive demonstrative pronoun *diese* in (1).

(1)

Diese	nicht	neue	Erkenntnis	konnte
this	not	new	insight	could
der	Markt	der	Möglichkeiten	am
the	market	of.the	possibilities	on.the
Sonnabend	in	Treuenbrietzen	bestens	
Saturday	in	Treuenbrietzen	in.the.best.way	
unterstreichen	.			
underline				

'The 'Market of Possibilities', held this Saturday in Treuenbrietzen, provided best evidence for this well-known (lit. 'not new') insight.' (PCC, #4794)

The phrase *diese nicht neue Erkenntnis* poses two challenges. First, it has to be recognized that the demonstrative pronoun is attributive, although it is separated from adjective and noun by *nicht* 'not'. Second, the phrase is in accusative case, although the morphology is ambiguous between accusative and nominative, and nominative case would be expected for a sentence-initial NP.

The Connexor analysis (Tapanainen and Järvinen, 1997) actually fails in both aspects (2).

(2) `PRON Dem FEM SG NOM` (Connexor)

The ontological analysis of this annotation begins by identifying the set of individuals from the Connexor Annotation Model that match it according to their `hasTag` (etc.) properties. The RDF triplet `connexor:NOM connexor:hasTagContaining 'NOM'`[4] indicates that the tag is an application of the individual `connexor:NOM`, an instance of `connexor:Case`. Further, the annotation matches `connexor:PRON` (an instance of `connexor:Pronoun`), etc. The result is a set of individuals that express different aspects of the meaning of the annotation.

For these individuals, the Annotation Model specifies superclasses (`rdf:type`) and other properties, i.e., `connexor:NOM connexor:hasCase connexor:NOM`, etc. The linguistic unit represented by the actual token can now be characterized by these properties: Every property applicable to a member in the individual set is assumed to be applicable to the linguistic unit as well. In order to save space, we use a notation closer to predicate logic (with the token as implicit subject). In terms of the Annotation Model, the token *diese* is thus described by the following descriptions:

[4]RDF triplets are quoted in simplified form, with XML namespaces replacing the actual URIs.

(3) `rdf:type(connexor:Pronoun)`
`connexor:hasCase(connexor:NOM) ...`

The Linking Model `connexor-link.rdf` provides us with the information that (i) `connexor:Pronoun` is a subclass of the Reference Model concept `olia:Pronoun`, (ii) `connexor:NOM` is an instance of the Reference Model concept `olia:Nominative`, and (iii) `olia:hasCase` is a subproperty of `olia:hasCase`.

Accordingly, the predicates that describe the token *diese* can be reformulated in terms of the Reference Model. `rdf:type(connexor:Pronoun)` entails `rdf:type(olia:Pronoun)`, etc. Similarly, we know that for some i:`olia:Nominative` it is true that `olia:hasCase(`i`)`, abbreviated here as `olia:hasCase(some olia:Nominative)`.

In this way, the grammatical information conveyed in the original Connexor annotation can be represented in an annotation-independent and tagset-neutral way as shown for the Connexor analysis in (4).

(4) `rdf:type(olia:PronounOrDeterminer)`
`rdf:type(olia:Pronoun)`
`olia:hasNumber(some olia:Singular)`
`olia:hasGender(some olia:Feminine)`
`rdf:type(olia:DemonstrativePronoun)`
`olia:hasCase(some olia:Nominative)`

Analogously, the corresponding RFTagger analysis (Schmid and Laws, 2008) given in (5) can be transformed into a description in terms of the OLiA Reference Model such as in (6).

(5) `PRO.Dem.Attr.-3.Acc.Sg.Fem` (RFTagger)

(6) `rdf:type(olia:PronounOrDeterminer)`
`olia:hasNumber(some olia:Singular)`
`olia:hasGender(some olia:Feminine)`
`olia:hasCase(some olia:Accusative)`
`rdf:type(olia:DemonstrativeDeterminer)`
`rdf:type(olia:Determiner)`

For every description obtained from these (and further) analyses, an integrated and consistent generalization can be established as described in the following section.

3 Processing linguistic annotations

3.1 Evaluation setup

Fig. 6 sketches the architecture of the evaluation environment set up for this study.[5] The input to the system is a set of documents with

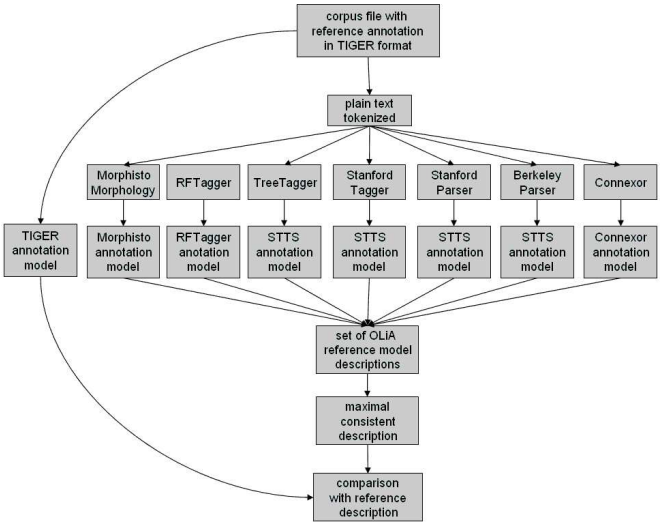

Figure 6: Evaluation setup

TIGER/NEGRA-style morphosyntactic or morphological annotation (Skut et al., 1998; Brants and Hansen, 2002) whose annotations are used as gold standard.

From the annotated document, the plain tokenized text is extracted and analyzed by one or more of the following NLP tools:

(i) Morphisto, a morphological analyzer without contextual disambiguation (Zielinski and Simon, 2008),

(ii) two part of speech taggers: the TreeTagger (Schmid, 1994) and the Stanford Tagger (Toutanova et al., 2003),

(iii) the RFTagger that performs part of speech and morphological analysis (Schmid and Laws, 2008),

(iv) two PCFG parsers: the StanfordParser (Klein and Manning, 2003) and the BerkeleyParser (Petrov and Klein, 2007), and

(v) the Connexor dependency parser (Tapanainen and Järvinen, 1997).

These tools annotate parts of speech, and those in (i), (iii) and (v) also provide morphological features. All components ran in parallel threads on the same machine, with the exception of Morphisto that was addressed as a web service. The set of matching Annotation Model individuals for every annotation and the respective set of Reference Model descriptions are determined by means of

[5] The code used for the evaluation setup is available under `http://multiparse.sourceforge.net`.

OLiA description	$\sum$	Morphisto	Connexor	RF Tagger	Tree Tagger	Stanford Tagger	Stanford Parser	Berkeley Parser
word class `type(...)`								
`PronounOrDeterminer`	7	1(4/4)*	1	1	1	1	1	1
`Determiner`	5.5	0.5**	0	1	1	1	1	1
`DemonstrativeDeterminer`	5.5	0.5**	0	1	1	1	1	1
`Pronoun`	1.5	0.5**	1	0	0	0	0	0
`DemonstrativePronoun`	1.5	0.5**	1	0	0	0	0	0
morphology `hasXY(...)`					n/a	n/a	n/a	n/a
`hasNumber(some Singular)`	2.5	0.5 (2/4)	1	1				
`hasGender(some Feminine)`	2.5	0.5 (2/4)	1	1				
`hasCase(some Accusative)`	1.5	0.5 (2/4)	0	1				
`hasCase(some Nominative)`	1.5	0.5 (2/4)	1	0				
`hasNumber(some Plural)`	0.5	0.5 (2/4)	0	0				

* Morphisto produces four alternative candidate analyses for this example, so every alternative analysis receives the confidence score 0.25

** Morphisto does not distinguish attributive and substitutive pronouns, it predicts `type(Determiner ⊔ Pronoun)`

Table 1: Confidence scores for *diese* in ex. (1)

the Pellet reasoner (Sirin et al., 2007) as described above.

A disambiguation routine (see below) then determines the maximal consistent set of ontological descriptions. Finally, the outcome of this process is compared to the set of descriptions corresponding to the original annotation in the corpus.

3.2 Disambiguation

Returning to examples (4) and (6) above, we see that the resulting set of descriptions conveys properties that are obviously contradicting, e.g., `hasCase(some Nominative)` besides `hasCase(some Accusative)`.

Our approach to disambiguation combines ontological consistency criteria with a confidence ranking. As we simulate an uninformed approach, the confidence ranking follows a majority vote.

For *diese* in (1), the consultation of all seven tools results a confidence ranking as shown in Tab. 1: If a tool supports a description with its analysis, the confidence score is increased by 1 (or by $1/n$ if the tool proposes n alternative annotations). A maximal consistent set of descriptions is then established as follows:

(i) Given a confidence-ranked list of available descriptions $S = (s_1, ..., s_n)$ and a result set $T = \emptyset$.

(ii) Let s_1 be the first element of $S = (s_1, ..., s_n)$.

(iii) If s_1 is consistent with every description $t \in T$, then add s_1 to T: $T := T \cup \{s_1\}$

(iv) Remove s_1 from S and iterate in (ii) until S is empty.

The consistency of ontological descriptions is defined here as follows:[6]

- Two concepts A and B are consistent iff

  $$A \equiv B \text{ or } A \sqsubseteq B \text{ or } B \sqsubseteq A$$

  Otherwise, A and B are disjoint.

- Two descriptions $pred_1(A)$ and $pred_2(B)$ are consistent iff

  A and B are consistent or
  $pred_1$ is neither a subproperty
  nor a superproperty of $pred_2$

This heuristic formalizes an implicit disjointness assumption for all concepts in the ontology (all concepts are disjoint unless one is a subconcept of the other). Further, it imposes an implicit cardinality constraint on properties (e.g., `hasCase(some Accusative)` and `hasCase(some Nominative)` are inconsistent because `Accusative` and `Nominative` are sibling concepts and thus disjoint).

For the example *diese*, the descriptions `type(Pronoun)` and `type(DemonstrativePronoun)` are inconsistent with `type(Determiner)`, and `hasNumber(some Plural)` is inconsistent with `hasNumber(some Singular)` (Figs. 2 and 4); these descriptions are thus ruled out. The `hasCase` descriptions have identical confidence scores, so that the first `hasCase` description that the algorithm encounters is chosen for the set of resulting descriptions, the other one is ruled out because of their inconsistency.

[6]The OLiA Reference Model does not specify disjointness constraints, and neither do GOLD or the DCR as External Reference Models. The axioms of the OntoTag ontologies, however, are specific to Spanish and cannot be directly applied to German.

	PCC	TIGER	NEGRA
best-performing tool (StanfordTagger)			
	.960	.956	.990*
average (and std. deviation) for tool combinations			
1 tool	.868 (.109)	.864 (.122)	.870 (.113)
2 tools	.928 (.018)	.931 (.021)	.943 (.028)
3 tools	.947 (.014)	.948 (.013)	.956 (.018)
4 tools	.956 (.006)	.955 (.009)	.963 (.013)
5 tools	.959 (.006)	.960 (.007)	.964 (.009)
6 tools	.963 (.003)	.963 (.007)	.965 (.007)
all tools	.967	.960	.965

* The Stanford Tagger was trained on the NEGRA corpus.

Table 2: Recall for `rdf:type` descriptions for word classes

	TIGER	NEGRA
1 tool	.678 (.106)	.660 (.091)
Morphisto	.573	.568
Connexor	.674	.662
RFTagger	.786	.751
2 tools	.761 (.019)	.740 (.012)
C+M	.738	.730
M+R	.769	.737
C+R	.773	.753
all tools	.791	.770

Table 3: Recall for morphological `hasXY()` descriptions

The resulting, maximal consistent set of descriptions is then compared with the ontological descriptions that correspond to the original annotation in the corpus.

4 Evaluation

Six experiments were conducted with the goal to evaluate the prediction of word classes and morphological features on parts of three corpora of German newspaper articles: NEGRA (Skut et al., 1998), TIGER (Brants et al., 2002), and the Potsdam Commentary Corpus (Stede, 2004, PCC). From every corpus 10,000 tokens were considered for the analysis.

TIGER and NEGRA are well-known resources that also influenced the design of several of the tools considered. For this reason, the PCC was consulted, a small collection of newspaper commentaries, 30,000 tokens in total, annotated with TIGER-style parts of speech and syntax (by members of the TIGER project). None of the tools considered here were trained on this data, so that it provides independent test data.

The ontological descriptions were evaluated for recall:[7]

$$(7)\quad recall(T) = \frac{\sum_{i=1}^{n} |D_{predicted}(t_i) \cap D_{target}(t_i)|}{\sum_{i=1}^{n} |D_{target}(t_i)|}$$

In (7), T is a text (a list of tokens) with $T = (t_1, ..., t_n)$, $D_{predicted}(t)$ are descriptions retrieved from the NLP analyses of the token t, and $D_{target}(t)$ is the set of descriptions that correspond to the original annotation of t in the corpus.

[7] Precision and accuracy may not be appropriate measurements in this case: Annotation schemes differ in their expressiveness, so that a description predicted by an NLP tool but not found in the reference annotation may nevertheless be correct. The RFTagger, for example, assigns demonstrative pronouns the feature '3rd person', that is not found in TIGER/NEGRA-style annotation because of its redundancy.

4.1 Word classes

Table 2 shows that the recall of `rdf:type` descriptions (for word classes) increases continuously with the number of NLP tools applied. The combination of all seven tools actually shows a better recall than the best-performing single NLP tool. (The NEGRA corpus is an apparent exception only; the exceptionally high recall of the Stanford Tagger reflects the fact that it was trained on NEGRA.)

A particularly high increase in recall occurs when tools are combined that compensate for their respective deficits. Morphisto, for example, generates alternative morphological analyses, so that the disambiguation algorithm performs a random choice between these. Morphisto has thus the worst recall among all tools considered (PCC .69, TIGER .65, NEGRA .70 for word classes). As compared to this, Connexor performs a contextual disambiguation; its recall is, however, limited by its coarse-grained word classes (PCC .73, TIGER .72, NEGRA .73). The combination of both tools yields a more detailed and context-sensitive analysis and thus results in a boost in recall by more than 13% (PCC .87, TIGER .86, NEGRA .86).

4.2 Morphological features

For morphological features, Tab. 3 shows the same tendencies that were also observed for word classes: The more tools are combined, the greater the recall of the generated descriptions, and the recall of combined tools often outperforms the recall of individual tools.

The three tools that provide morphological annotations (Morphisto, Connexor, RFTagger) were evaluated against 10,000 tokens from TIGER and NEGRA respectively. The best-performing tool was the RFTagger, which possibly reflects the fact

that it was trained on TIGER-style annotations, whereas Morphisto and Connexor were developed on the basis of independent resources and thus differ from the reference annotation in their respective degree of granularity.

5 Summary and Discussion

With the ontology-based approach described in this paper, the performance of annotation tools can be evaluated on a conceptual basis rather than by means of a string comparison with target annotations. A formal model of linguistic concepts is extensible, finer-grained and, thus, potentially more adequate for the integration of linguistic annotations than string-based representations, especially for heterogeneous annotations, if the tagsets involved are structured according to different design principles (e.g., due to different terminological traditions, different communities involved, etc.).

It has been shown that by **abstracting from tool-specific representations** of linguistic annotations, annotations from different tagsets can be represented with reference to the OLiA ontologies (and/or with other OWL/RDF-based terminology repositories linked as External Reference Models). In particular, it is possible to compare an existing reference annotation with annotations produced by NLP tools that use independently developed and differently structured annotation schemes (such as Connexor vs. RFTagger vs. Morphisto).

Further, an algorithm for the **integration of different annotations** has been proposed that makes use of a majority-based confidence ranking and ontological consistency conditions. As consistency conditions are not formally defined in the OLiA Reference Model (which is expected to inherit such constraints from External Reference Models), a heuristic, structure-based definition of consistency was applied.

This heuristic consistency definition is overly rigid and rules out a number of consistent alternative analyses, as it is the case for overlapping categories.[8] Despite this rigidity, we witness an **increase of recall** when multiple alternative analyses are integrated. This increase of recall may result from a compensation of tool-specific deficits, e.g., with respect to annotation granularity. Also, the improved recall can be explained by a compensation of overfitting, or deficits that are inherent to a particular approach (e.g., differences in the coverage of the linguistic context).

It can thus be stated that the integration of multiple alternative analyses has the potential to produce linguistic analyses that are both more robust and more detailed than those of the original tools.

The primary field of application of this approach is most likely to be seen in a context where applications are designed that make direct use of OWL/RDF representations as described, for example, by Hellmann (2010). It is, however, also possible to use ontological representations to bootstrap novel and more detailed annotation schemes, cf. Zavrel and Daelemans (2000). Further, the conversion from string-based representations to ontological descriptions is reversible, so that results of ontology-based disambiguation and validation can also be reintegrated with the original annotation scheme. The idea of such a reversion algorithm was sketched by Buyko et al. (2008) where the OLiA ontologies were suggested as a means to translate between different annotation schemes.[9]

6 Extensions and Related Research

Natural extensions of the approach described in this paper include:

(i) Experiments with formally defined consistency conditions (e.g., with respect to restrictions on the domain of properties).

(ii) Context-sensitive disambiguation of morphological features (e.g., by combination with a chunker and adjustment of confidence scores for morphological features over all tokens in the current chunk, cf. Kermes and Evert, 2002).

(iii) Replacement of majority vote by more elaborate strategies to merge grammatical analyses.

[8] Preposition-determiner compounds like German *am* 'on the', for example, are both prepositions and determiners.

[9] The mapping from ontological descriptions to tags of a particular scheme is possible, but neither trivial nor necessarily lossless: Information of ontological descriptions that cannot be expressed in the annotation scheme under consideration (e.g., the distinction between attributive and substitutive pronouns in the Morphisto scheme) will be missing in the resulting string representation. For complex annotations, where ontological descriptions correspond to different substrings, an additional 'tag grammar' may be necessary to determine the appropriate ordering of substrings according to the annotation scheme (e.g., in the Connexor analysis).

(iv) Application of the algorithm for the ontological processing of node labels and edge labels in syntax annotations.

(v) Integration with other ontological knowledge sources in order to improve the recall of morphosyntactic and morphological analyses (e.g., for disambiguating grammatical case).

Extensions (iii) and (iv) are currently pursued in an ongoing research effort described by Chiarcos et al. (2010). Like morphosyntactic and morphological features, node and edge labels of syntactic trees are ontologically represented in several Annotation Models, the OLiA Reference Model, and External Reference Models, the merging algorithm as described above can thus be applied for syntax, as well. Syntactic annotations, however, involve the additional challenge to align different structures before node and edge labels can be addressed, an issue not further discussed here for reasons of space limitations.

Alternative strategies to merge grammatical analyses may include alternative voting strategies as discussed in literature on classifier combination, e.g., weighted majority vote, pairwise voting (Halteren et al., 1998), credibility profiles (Tufiş, 2000), or hand-crafted rules (Borin, 2000). A novel feature of our approach as compared to existing applications of these methods is that confidence scores are not attached to plain strings, but to ontological descriptions: Tufiş, for example, assigned confidence scores not to tools (as in a weighted majority vote), but rather, assessed the 'credibility' of a tool *with respect to the predicted tag*. If this approach is applied to ontological descriptions in place of tags, it allows us to consider the credibility of pieces of information regardless of the actual string representation of tags. For example, the credibility of `hasCase` descriptions can be assessed independently from the credibility of `hasGender` descriptions even if the original annotation merged both aspects in one single tag (as the RFTagger does, for example, cf. ex. 5).

Extension (v) has been addressed in previous research, although mostly with the opposite perspective: Already Cimiano and Reyle (2003) noted that the integration of grammatical and semantic analyses may be used to resolve ambiguity and underspecifications, and this insight has also motivated the ontological representation of linguistic resources such as WordNet (Gangemi et al., 2003), FrameNet (Scheffczyk et al., 2006), the linking of corpora with such ontologies (Hovy et al., 2006), the modelling of entire corpora in OWL/DL (Burchardt et al., 2008), and the extension of existing ontologies with ontological representations of selected linguistic features (Buitelaar et al., 2006; Davis et al., 2008).

Aguado de Cea et al. (2004) sketched an architecture for the closer ontology-based integration of grammatical and semantic information using OntoTag and several NLP tools for Spanish. Aguado de Cea et al. (2008) evaluate the benefits of this approach for the Spanish particle *se*, and conclude for this example that the combination of multiple tools yields more detailed and more accurate linguistic analyses of particularly problematic, polysemous function words. A similar increase in accuracy has also been repeatedly reported for ensemble combination approaches, that are, however, limited to tools that produce annotations according to the *same* tagset (Brill and Wu, 1998; Halteren et al., 2001).

These observations provide further support for our conclusion that the ontology-based integration of morphosyntactic analyses enhances both the robustness and the level of detail of morphosyntactic and morphological analyses. Our approach extends the philosophy of ensemble combination approaches to NLP tools that do not only employ different strategies and philosophies, but also different annotation schemes.

Acknowledgements

From 2005 to 2008, the research on linguistic ontologies described in this paper was funded by the German Research Foundation (DFG) in the context of the Collaborative Research Center (SFB) 441 "Linguistic Data Structures", Project C2 "Sustainability of Linguistic Resources" (University of Tübingen), and since 2007 in the context of the SFB 632 "Information Structure", Project D1 "Linguistic Database" (University of Potsdam). The author would also like to thank Julia Ritz, Angela Lahee, Olga Chiarcos and three anonymous reviewers for helpful hints and comments.

References

G. Aguado de Cea, Á. I. de Mon-Rego, A. Pareja-Lora, and R. Plaza-Arteche. 2002. OntoTag: A semantic web page linguistic annotation model. In *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation and Knowledge Markup*, Lyon, France, July.

G. Aguado de Cea, A. Gomez-Perez, I. Alvarez de Mon, and A. Pareja-Lora. 2004. OntoTag's linguistic ontologies: Improving semantic web annotations for a better language understanding in machines. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*, Las Vegas, Nevada, USA, April.

G. Aguado de Cea, J. Puch, and J.Á. Ramos. 2008. Tagging Spanish texts: The problem of "se". In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

A. Aschenbrenner, P. Gietz, M.W. Küster, C. Ludwig, and H. Neuroth. 2006. TextGrid. A modular platform for collaborative textual editing. In *Proceedings of the International Workshop on Digital Library Goes e-Science (DLSci06)*, pages 27–36, Alicante, Spain, September.

D. Bakker, O. Dahl, M. Haspelmath, M. Koptjevskaja-Tamm, C. Lehmann, and A. Siewierska. 1993. EUROTYP guidelines. Technical report, European Science Foundation Programme in Language Typology.

B. Bickel and J. Nichols. 2000. The goals and principles of AUTOTYP. `http://www.uni-leipzig.de/~autotyp/theory.html`. version of 01/12/2007.

B. Bickel and J. Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the LREC 2002 Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain, May.

L. Borin. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, May, 31st – June, 2nd.

S. Brants and S. Hansen. 2002. Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1643–1649, Las Palmas, Spain, May.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol, Bulgaria, September.

E. Brill and J. Wu. 1998. Classifier combination for improved lexical disambiguation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 191–195, Montréal, Canada, August.

P. Buitelaar, T. Declerck, A. Frank, S. Racioppa, M. Kiesel, M. Sintek, R. Engel, M. Romanelli, D. Sonntag, B. Loos, V. Micelli, R. Porzel, and P. Cimiano. 2006. LingInfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May.

A. Burchardt, S. Padó, D. Spohr, A. Frank, and U. Heid. 2008. Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In *Proceedings of the 3rd International Joint Conference on NLP (IJCNLP 2008)*, Hyderabad, India, January.

E. Buyko, C. Chiarcos, and A. Pareja-Lora. 2008. Ontology-based interface specifications for a NLP pipeline architecture. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

M. Carl, C. Pease, L.L. Iomdin, and O. Streiter. 2000. Towards a dynamic linkage of example-based and rule-based machine translation. *Machine Translation*, 15(3):223–257.

C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede. 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues*, 49(2).

C. Chiarcos, K. Eckart, and J. Ritz. 2010. Creating and exploiting a resource of parallel parses. In *4th Linguistic Annotation Workshop (LAW 2010), held in conjunction with ACL-2010*, Uppsala, Sweden, July.

C. Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16. Foundations of Ontologies in Text Technology, Part II: Applications.

C. Chiarcos. 2010. Grounding an ontology of linguistic annotations in the Data Category Registry. In *Workshop on Language Resource and Language Technology Standards (LR&LTS 2010), held in conjunction with LREC 2010*, Valetta, Malta, May.

P. Cimiano and U. Reyle. 2003. Ontology-based semantic construction, underspecification and disambiguation. In *Proceedings of the Lorraine/Saarland Workshop on Prospects and Recent Advances in the Syntax-Semantics Interface*, pages 33–38, Nancy, France, October.

B. Crysmann, A. Frank, B. Kiefer, S. Müller, G. Neumann, J. Piskorski, U. Schäfer, M. Siegel, H. Uszkoreit, F. Xu, M. Becker, and H. Krieger. 2002. An

integrated architecture for shallow and deep processing. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 441–448, Philadelphia, Pennsylvania, USA, July.

B. Davis, S. Handschuh, A. Troussov, J. Judge, and M. Sogrin. 2008. Linguistically light lexical extensions for ontologies. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

S. Dipper, M. Götze, and S. Skopeteas, editors. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*. Interdisciplinary Studies on Information Structure (ISIS), Working Papers of the SFB 632; 7. Universitätsverlag Potsdam, Potsdam, Germany.

M.T. Egner, M. Lorch, and E. Biddle. 2007. UIMA Grid: Distributed large-scale text analysis. In *Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGRID'07)*, pages 317–326, Rio de Janeiro, Brazil, May.

S. Farrar and D.T. Langendoen. 2003. Markup and the GOLD ontology. In *EMELD Workshop on Digitizing and Annotating Text and Field Recordings*. Michigan State University, July.

A. Gangemi, R. Navigli, and P. Velardi. 2003. The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In R. Meersman and Z. Tari, editors, *Proceedings of On the Move to Meaningful Internet Systems (OTM2003)*, pages 820–838, Catania, Italy, November.

P. Gietz, A. Aschenbrenner, S. Budenbender, F. Jannidis, M.W. Küster, C. Ludwig, W. Pempe, T. Vitt, W. Wegstein, and A. Zielinski. 2006. TextGrid and eHumanities. In *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (E-SCIENCE '06)*, pages 133–141, Amsterdam, The Netherlands, December.

H. van Halteren, J. Zavrel, and W. Daelmans. 1998. Improving data driven wordclass tagging by system combination. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, Montréal, Canada, August.

H. van Halteren, J. Zavrel, and W. Daelmans. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics*, 27(2):199–229.

S. Hellmann. 2010. The semantic gap of formalized meaning. In *The 7th Extended Semantic Web Conference (ESWC 2010)*, Heraklion, Greece, May 30th – June 3rd.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2006)*, pages 57–60, New York, June.

N. Ide and L. Romary. 2004. A registry of standard data categories for linguistic annotation. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, pages 135–39, Lisboa, Portugal, May.

M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S.E. Wright. 2009. ISOcat: remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261–276.

H. Kermes and S. Evert. 2002. YAC – A recursive chunker for unrestricted German text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1805–1812, Las Palmas, Spain, May.

J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus – A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July.

G. Leech and A. Wilson. 1996. EAGLES recommendations for the morphosyntactic annotation of corpora. Version of March 1996.

T. Luís and D.M. de Matos. 2009. High-performance high-volume layered corpora annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW-III) held in conjunction with ACL-IJCNLP 2009*, pages 99–107, Singapore, August.

M. Mandel. 2006. Integrated annotation of biomedical text: Creating the PennBioIE corpus. In *Text Mining Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK, March.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

R. Meyer. 2003. Halbautomatische morphosyntaktische Annotation russischer Texte. In R. Hammel and L. Geist, editors, *Linguistische Beiträge zur Slavistik aus Deutschland und Österreich. X. JungslavistInnen-Treffen, Berlin 2001*, pages 92–105. Sagner, München.

S. Petrov and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2007)*, pages 404–411, Rochester, NY, April.

S. Petrova, C. Chiarcos, J. Ritz, M. Solf, and A. Zeldes. 2009. Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. *Traitement automatique des langues et langues anciennes*, 50(2):47–71.

G. Rehm, R. Eckart, and C. Chiarcos. 2007. An OWL- and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria, September.

G. Sampson. 1995. *English for the computer: The SUSANNE corpus and analytic scheme*. Oxford University Press.

A. Saulwick, M. Windhouwer, A. Dimitriadis, and R. Goedemans. 2005. Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proceedings of the 17th Conference on Advanced Information Systems Engineering (CAiSE'05)*, Porto, Portugal, June.

J. Scheffczyk, A. Pease, and M. Ellsworth. 2006. Linking FrameNet to the suggested upper merged ontology. In *Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS 2006)*, pages 289–300, Baltimore, Maryland, USA, November.

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

H. Schmid and F. Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK, August.

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

T. Schmidt, C. Chiarcos, T. Lehmberg, G. Rehm, A. Witt, and E. Hinrichs. 2006. Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In *Proceedings of the E-MELD workshop on Digital Language Documentation: Tools and Standards: The State of the Art*, East Lansing, Michigan, US, June.

S. Sharoff, M. Kopotev, T. Erjavec, A. Feldman, and D. Divjak. 2008. Designing and evaluating Russian tagsets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, and Y. Katz. 2007. Pellet: A practical OWL/DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53.

W. Skut, T. Brants, B. Krenn, and H. Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *In Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany, August.

M. Stede. 2004. The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, July.

P. Tapanainen and T. Järvinen. 1997. A nonprojective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington, DC, April.

K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003)*, Edmonton, Canada, May.

D. Tufiş. 2000. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1105–1112, Athens, Greece, May, 31st – June, 2nd.

A. Wagner and B. Zeisler. 2004. A syntactically annotated corpus of Tibetan. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal, May.

J. Zavrel and W. Daelemans. 2000. Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, May, 31st – June, 2nd.

A. Zielinski and C. Simon. 2008. Morphisto: An open-source morphological analyzer for German. In *Proceedings of the Conference on Finite State Methods in Natural Language Processing (FSMNLP)*, Ispra, Italy, September.