OLiA – Ontologies of Linguistic Annotation

Christian Chiarcos
Information Sciences Institute, University of Southern California, chiarcos@isi.edu

Abstract. This paper describes the Ontologies of Linguistic Annotation (OLiA) as one of the data sets currently available as part of Linguistic Linked Open Data (LLOD) cloud. The OLiA ontologies represent a repository of annotation terminology for various linguistic phenomena on a great band-width of languages, they have been used to facilitate interoperability and information integration of linguistic annotations in corpora, NLP pipelines, and lexical-semantic resources.

Keywords: linguistic terminology, grammatical categories, annotation schemes, corpora, NLP, tag sets, annotation schemes

1. Background

The heterogeneity of linguistic annotations has been recognized as a key problem limiting the interoperability and reusability of NLP tools and linguistic data collections. Several repositories of linguistic annotation terminology have been developed to facilitate annotation interoperability by means of a joint level of representation, or an 'interlingua', the most prominent probably being the General Ontology of Linguistic Description [13, GOLD] and the ISO TC37/SC4 Data Category Registry [18, ISOcat].

Still, these repositories are developed by different communities, and are thus not always compatible with each other, neither with respect to their definitions, or their technologies (e.g., there is no commonly agreed formalism to link linguistic annotations to terminology repositories), and harmonization efforts are still in their early stages [17].

The Ontologies of Linguistic Annotation (OLiA) have been developed to facilitate the development of applications that take benefit of a well-defined terminological backbone even before the GOLD and ISOcat repositories have converged into a generally accepted reference terminology: They introduce an intermediate level of representation between ISOcat, GOLD and other repositories of linguistic reference terminology and are interconnected with these resources, and they provide not only means to formalize reference categories, but also annotation schemes, and the way that these are linked with reference categories.

2. Architecture

The **Ontologies of Linguistic Annotations** [3] represent a modular architecture of OWL/DL ontologies that formalize several intermediate steps of the mapping between annotations, a 'Reference Model' and existing terminology repositories ('External Reference Models').

The OLiA ontologies were developed as part of an infrastructure for the sustainable maintenance of linguistic resources [27], and their primary fields of application include the formalization of annotation schemes and concept-based querying over heterogeneously annotated corpora [23,9].

In the OLiA architecture, four different types of ontologies are distinguished:

- The OLIA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is derived from existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.
- Multiple OLIA ANNOTATION MODELS formalize annotation schemes and tagsets. Annotation Models are based on the original documentation, so that they provide an interpretation-independent representation of the annotation scheme.
- For every Annotation Model, a LINKING MODEL defines $\sqsubseteq$ 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.

- Existing terminology repositories can be integrated as EXTERNAL REFERENCE MODELS, if they are represented in OWL/DL. Then, Linking Models specify $\sqsubseteq$ relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g., `olia:Determiner`) and grammatical features (e.g., `olia:Accusative`), as well as properties that define relations between these (e.g., `olia:hasCase`). Far from being yet another annotation terminology ontology, the OLiA Reference Model does not introduce its own view on the linguistic world, but rather, it is a derivative of EAGLES [19], MULTEXT/East [12], and GOLD [13] that was introduced as a technical means to interpret linguistic annotations with respect to these terminological repositories, and further enriched with information drawn from the annotation schemes it was applied to.

Conceptually, Annotation Models differ from the Reference Model in that they include not only concepts and properties, but also individuals: Individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model.

3. Data Set Description

The OLiA ontologies are available from `http://purl.org/olia` under a Creative Commons Attribution license (CC-BY).

The OLiA ontologies cover different grammatical phenomena, including inflectional morphology, word classes, phrase and edge labels of different syntax annotations, as well as prototypes for discourse annotations (coreference, discourse relations, discourse structure and information structure). Annotations for lexical semantics are only covered to the extent that they are encoded in syntactic and morphosyntactic annotation schemes. For lexical semantic annotations in general, a number of reference resources is already available, including RDF versions of WordNet and FrameNet.

In recent years, the OLiA ontologies have been substantially extended. At the time of writing, the OLiA Reference Model distinguishes 14 `MorphologicalCategorys` (morphemes), 263 `MorphosyntacticCategorys` (word classes), 83 `SyntacticCategorys` (phrase labels), and 326 different values for 16 `MorphosyntacticFeatures`, 4 `MorphosyntacticFeatures`, 4 `SyntacticFeatures` and 4 `SemanticFeatures` (for glosses, part-of-speech annotation and for edge labels in syntax annotation).

As for morphological, morphosyntactic and syntactic annotations, the OLiA ontologies include 32 Annotation Models for about 70 different languages, including several multi-lingual annotation schemes, e.g., EAGLES [3] for 11 Western European languages, and MULTEXT/East [7] for 15 (mostly) Eastern European languages. As for non-(Indo-)European languages, the OLiA ontologies include morphosyntactic annotation schemes for languages of the Indian subcontinent, for Arabic, Basque, Chinese, Estonian, Finnish, Hausa, Hungarian and Turkish, as well as multi-lingual schemes applied to languages of Africa, the Americas, the Pacific and Australia. The OLiA ontologies also cover historical language stages, including Old High German, Old Norse and Old/Classical Tibetan. Additionally, 7 Annotation Models for different resources with discourse annotations have been developed.

External reference models currently linked to the OLiA Reference Model include GOLD [3], the OntoTag ontologies [1], an ontological remodeling of ISOcat [4], and the Typological Database System (TDS) ontologies [25]. Unfortunately, neither of these external reference models can currently be redistributed because of an uncertain licensing situation (GOLD and ISOcat lack explicit license information)[1] or other restrictions (the OntoTag and TDS ontologies are available to the author but have not been publicly released).

In this context, the function of the OLiA Reference Model is not to provide a novel and independent view on linguistic terminology, but rather to serve as a stable intermediate representation between (ontological models of) annotation schemes and these terminology repositories. This allows any concept that can be expressed in terms of the OLiA Reference Model also to be interpreted in the context of ISOcat, GOLD, OntoTag or TDS.

As compared to a direct linking between annotation models and these terminology repositories, the modular structure limits the number of linkings that need to be defined (if a new Annotation Model is linked to

[1] Nevertheless, the developers are sympathetic to the idea of releasing this data under an open license, Helen Aristar-Dry (for GOLD) and Menzo Windhouwer (for ISOcat), pers. communication, June 2012.

the Reference Model, it inherits its linking with ISOcat, GOLD, OntoTag and TDS), and also, it provides stability (GOLD and ISOcat are developed in community processes with occasional revisions), a clear and non-redundant taxonomical organization (similar to GOLD, TDS and OntoTag, but very different from the semi-structured ISOcat) and establishes interoperability between GOLD and ISOcat (that – despite ongoing harmonization efforts [17] – are maintained by different communities and developed independently). Using the OLiA Reference Model, it is thus possible to develop applications that are interoperable in terms of GOLD *and* ISOcat even though both are still under development and both differ in their conceptualizations. Such applications are briefly described in the following section.

4. Application

Initially, the OLiA ontologies have been intended to serve a **documentation function**, i.e., as a formal means to specify the semantics of annotation schemes [27]. From the ontologies, dynamic HTML can be generated,[2] and tags in the annotation can be represented as hyperlinks pointing to the corresponding definition [9].

In earlier **corpus query systems**, e.g., ANNIS [8], and SPLICR [23], OLiA was used to formulate interoperable corpus queries: Instead of querying for `cat="NX"` to retrieve noun phrases from the TüBa-D/Z corpus [31] or `cat="NP"` on the NEGRA corpus [29, both are corpora of German newspaper text], a query for `cat in {olia:NounPhrase}` was expanded into a disjunction of possible tags [8]. If corpora are represented as Linked Data, they can be directly linked with OLiA Annotation Models, and queried with SPARQL without a query preprocessor [5].

In a similar vein, OLiA can be employed in **NLP pipeline systems** and other NLP pipeline systems for tagset-independent, interoperable information processing [1]. In this function, OLiA is part of the specifications of the NLP Interchange Format (NIF).[3]

Figure 1 illustrates how annotations can be mapped onto Reference Model concepts for the German phrase *Diese nicht neue Erkenntnis* 'this well-known (lit. not new) insight' from the Potsdam Commentary Corpus [30, file 4794], with part-of-speech annotations according to the STTS scheme [26]: The tag `PDAT` matches the surface string of the individual `stts:PDAT` from the STTS Annotation Model.[4] The superconcept `stts:AttributiveDemonstrativePronoun` is a subconcept of `olia:DemonstrativeDeterminer` (STTS Linking Model).[5] The word *diese* 'this' from the example can thus be described in terms of the OLiA Reference Model as `olia:DemonstrativeDeterminer`, etc.

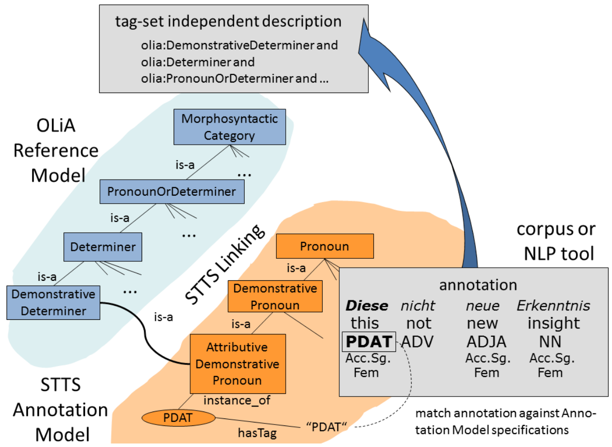

Fig. 1. Interpreting annotations in terms of the OLiA Reference Model

These ontology-based descriptions are comparable across different corpora and/or NLP tools, across different languages, and even across different types of language resources: Recently, the OLiA ontologies have also been applied to represent grammatical specifications of machine-readable dictionaries, that are thus interoperable with OLiA-linked corpora [20]. Moreover, through the linking with External Reference Models like GOLD and ISOcat, OLiA-linked resources are also interoperable with resources directly grounded in either GOLD or ISOcat.

Using Semantic Web formalisms to represent corpora and annotations also provides us with the possibility to develop novel, **ontology-based NLP** algorithms. One application are ensemble combination architectures, where different NLP modules (say, part-of-speech taggers) are applied in parallel, so that they produce annotations for one particular phenomenon, and that these annotations are then integrated. Using

[2] `http://code.google.com/p/co-ode-owl-plugins/wiki/OWLDoc`

[3] `http://nlp2rdf.org/nif-1-0`

[4] `http://purl.org/olia/stts.owl`

[5] `http://purl.org/olia/stts-link.rdf`

OLiA Reference Model specifications to integrate the analyses of multiple NLP tools for German, [10] could show that a simple majority-based combination could increase both the robustness and the level of detail of morphosyntactic and morphological analyses. Similar results have been obtained with the OntoTag ontologies for Spanish [22].

5. Discussion

This paper summarized the development of the OLiA ontologies since 2006, their current status, and a number of applications that have been developed on this basis.

The fundamental idea of the OLiA architecture is that annotation schemes are linked to community-maintained terminology repositories through an intermediate 'Reference Model', thereby minimizing the number of mappings necessary establish interoperability of one annotation scheme with multiple terminology repositories. Further, annotation schemes and their linking to the Reference Model are formalized as separate OWL/DL ontologies, so that interpretation-independent conceptualization (annotation documentation) and its interpretation in terms of the Reference Model (linking) are properly distinguished.

The OLiA ontologies differ from related approaches in that they take a focus on modeling annotation schemes and their linking with reference categories rather than merely providing reference categories. The differentiation of Annotation Models, the OLiA Reference Model and External Reference Models (community-maintained terminology repositories) represents increasing levels of abstraction, and, possibly, loss of information. However, no information about the original annotation is lost, and tools may chose the appropriate level of abstraction. Unlike a direct mapping approach, OLiA allows to recover information about sources of mismatches between Reference Model concepts and Annotation Model concepts, because a declarative linking is provided that allows inspection and refinement using standard RDF/OWL tools.

The relationship between annotations and reference concept is not only represented in a transparent way, but also, conceptual *mis*matches can be represented. Many tagsets for part-of-speech annotation, for example, introduce hybrid categories to represent either conceptual overlap/fusion or ambiguity using OWL/DL constructs to represent conjunction ($\sqcap$) or disjunction ($\sqcup$). As compared to tagset-specific solutions [24, | for ambiguities, and + for cliticization/fusion], OWL/DL provides a W3C-standardized vocabulary to express these relationships, that also extends beyond individual tagsets. Another difference is that negation (owl:complementOf) is available in the linking. This is of particular importance for the linking between External Reference Models and the OLiA Reference Model. For example, an olia:ProQuantifier (pronominal quantifier, can substitute for an independent noun phrase, e.g., *someone*) can be defined as subclass of gold:Quantifier. According to its definition, however, gold:Quantifier primarily pertains to Determiners, so that a more appropriate superclass would be gold:Quantifier$\sqcap\neg$gold:Determiner.

The physical separation of Linking Models from Annotation Models and Reference Model introduces a clear distinction between externally provided information and the ontology engineer's interpretation. Annotation Models formalize annotation documentation, and the Reference Model is based on a generalization of a broad band-width of resources. However, there may be different terminological traditions involved, so that apparently similar concepts found in Reference Model and Annotation Model are in fact unrelated. If nevertheless an incorrect identification takes place, the linking can be inspected by standard ontology browsers, and corrected independently from the interpretation-invariant Annotation Model and Reference Model. Furthermore, *multiple* linkings between an Annotation Model and the Reference Model can be implemented, e.g., to accommodate for systematic tagger errors (i.e., more extensive usage of owl:join), or for multiple dialects of the same tagset (e.g., the STTS tagset distinguishes indefinite attributive pronouns in indefinite noun phrases [PIAT] and in definite noun phrases [PIDAT], but in the TüBa-D/Z corpus [31], PIAT covers both uses).

In ISOcat, the problem of conflicting interpretations of data categories is currently *not* addressed, and the definitions provided are not always sufficient to distinguish classes, e.g., the category definite/DC-2004 is defined as 'value referring to the capacity of identification of an entity'. The concept is (at least partially) grounded in MULTEXT/East [15], which, however, conflate different uses of 'definite': (1) postfixed Determiner in Romanian, Bulgarian and Persian nouns or adjectives, (2) difference between 'full' and 'reduced' adjectives in Slavic (diachronically, full forms reflect a clitic pronominal), (3) a pattern of quantifier agreement in Slavic, and (4) the so-called 'definite conjunction' of Hungarian verbs. Even though the generic definition captures most of these different meanings, they remain incompatible with each other. However, without modeling relations between different language-specific annotation schemes and a data category registry from a global perspective, it is possible that such ill-defined data categories and/or links remain *undetected*. Within MULTEXT/East, for example, only the ontological modeling of language-specific annotation schemes and the common morphosyntactic specifications led to the proper differentiation between these different conceptions of 'definite' [7]. The OLiA Reference Model provides such a fully developed taxonomy of linguistic categories. Recent activities to augment ISOcat with a relation category registry [28] may eventually lead to a comparable global perspective, so that the problem of conflicting interpretations of data categories may become more obvious to ISOcat developers, but these are still on-going developments.

In comparison to GOLD, OLiA is more focused on NLP and corpus interoperability, whereas GOLD originates from the language documentation community. Therefore, a number of data categories commonly assumed in NLP were not originally represented in GOLD. For example, `gold:CommonNoun` was added only recently (between 2006 and 2008), following a suggestion by the author. While the GOLD community process will eventually lead to a compensation of such coverage issues, a more fundamental problem is that the views of academic linguists and NLP engineers may deviate with respect to the overarching taxonomy of concepts. GOLD, for example, seems to conflate both semantic roles ('case' in the sense of [14], e.g., `gold:BenefactiveCase`) and syntactic roles under `gold:CaseProperty`. Therefore, OLiA adopts a relatively agnostic view on the taxonomical order of concepts. While the taxonomy is modeled in a specific way (mostly following established annotation schemes), it is not assumed that this way of modeling is the only possibility. In fact, alternative taxonomies can be formulated as External Reference Models, and OWL/DL-based allows to formulate specific conditions for the linking, including the use of negation and disjunction. Consequently, mismatches can be represented. (As opposed to this, GOLD Community of Practice Extensions are assumed to adopt the GOLD hierarchy and only to extend it, not to redefine it.)

Conceptually, the OLiA ontologies are closer related to the OntoTag ontologies [2], that were also applied to develop NLP applications on the basis of ontological representations of linguistic annotations [22]. One important difference is that the OntoTag ontologies are considering only the languages of the Iberian peninsula (in particular Spanish), that they are partially designed with a top-down perspective (whereas the development of the OLiA Reference Model is guided by the annotation schemes it is applied to) and are thus richer in consistency constraints (that are, however, often language-specific), and that the OntoTag ontologies are not publicly available at the moment. Within the OLiA architecture, the morphosyntactic layer of the OntoTag ontologies is integrated as an External Reference Model [1].

The OLiA ontologies may play an important role in NLP, corpus and annotation interoperability in that they relate these activities to initiatives in different linguistic communities to establish reference repositories for linguistic annotation terminology, e.g., recent developments towards the creation of a Linguistic Linked Open Data (LLOD) cloud. In this context, the OLiA ontologies are used to provide linguistic reference terminology for lexical-semantic resources such as *lemon* [21] and Uby [11] as well as for linguistic corpora such as the Manually Annotated Sub-Corpus of the Open American National Corpus [16].[6]

Acknowledgments

The OLiA ontologies have originally been developed at the Collaborative Research Center (SFB) 441 "Linguistic Data Structures" (University of Tübingen) in the context of the project "Sustainability of Linguistic Resources" in cooperation with SFB 632 "Information Structure" (University of Potsdam, Humboldt-University Berlin) and SFB 538 "Multilingualism" (University of Hamburg) from 2006 to 2008. From 2007 to 2011, they have been maintained and further developed in the context of SFB 632 in the context of the project "Linguistic Data Base". In 2012, the author continued his research on OLiA in the context of PostDoc fellowship at the Information Sciences Institute of the University of Southern California funded by the German Academic Exchange Service (DAAD).

[6] A Linked Data version of the corpus, MASC 1.0.3, was generated from data available under `http://datahub.io/dataset/masc` in preparation of the MLODE workshop. It is part of the LLOD cloud diagram, but not yet linked with other resources as it will soon be deprecated by the release of a new version of the corpus and a revision of its data model.

In parts, this data set description is based on [6], shortened, updated and thoroughly revised.

References

[1] Buyko E, Chiarcos C, Pareja-Lora A (2008) Ontology-based interface specifications for a NLP pipeline architecture. In: Proc. LREC 2008, Marrakech, Morocco

[2] Aguado de Cea G, Gomez-Perez A, Alvarez de Mon I, Pareja-Lora A (2004) OntoTag's linguistic ontologies. In: Proc. Information Technology: Coding and Computing (ITCC'04), Washington, DC, USA

[3] Chiarcos C (2008) An ontology of linguistic annotations. LDV Forum 23(1):1–16

[4] Chiarcos C (2010) Grounding an ontology of linguistic annotations in the Data Category Registry. In: LREC 2010 Workshop on Language Resource and Language Technology Standards (LT&LTS), Valetta, Malta, pp 37–40

[5] Chiarcos C (2012) Interoperability of Corpora and Annotations. In: Chiarcos C, Nordhoff S, Hellmann S (eds) Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata, Springer, Heidelberg, pp 161–179

[6] Chiarcos C (2012) Ontologies of linguistic annotation: Survey and perspectives. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp 303–310

[7] Chiarcos C, Erjavec T (2011) OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In: Proc. 5th Linguistic Annotation Workshop, held in conjunction with ACL-HTL 2011, Portland, pp 11–20

[8] Chiarcos C, Götze M (2007) A linguistic database with ontology-sensitive corpus querying. system demonstration at Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV 2007). T

[9] Chiarcos C, Dipper S, Götze M, et al (2008) A flexible framework for integrating annotations from different tools and tag sets. TAL (Traitement Automatique des Langues) 49(2)

[10] Chiarcos C, Hellmann S, Nordhoff S (2011) Towards a linguistic linked open data cloud: The open linguistics working group. Traitement automatique des langues 52(3):245–275

[11] Eckle-Kohler J, McCrae J, Chiarcos C (this vol.) *lemonUby* - a large, interlinked, syntactically-rich resource for ontologies. submitted to this volume

[12] Erjavec T (2004) MULTEXT-East version 3. In: Proc. LREC 2004, Lisboa, Portugal, pp 1535–1538

[13] Farrar S, Langendoen DT (2010) An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In: Witt AW, Metzing D (eds) Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology, Springer, Dordrecht

[14] Fillmore CJ (1968) The case for case. In: Bach E, Harms R (eds) Universals in Linguistic Theory, Holt, Rinehart, and Winston, New York, pp 1–88

[15] Francopoulo G, Declerck T, Sornlertlamvanich V, et al (2008) Data Category Registry: Morpho-syntactic and syntactic profiles. In: Proc. LREC-2008 Workshop on Uses and Usage of Language Resource-Related Standards, Marrakech, Morocco

[16] Ide N, Baker C, Fellbaum C, Fillmore C, Passonneau R (2008) MASC: The Manually Annotated Sub-Corpus of American English. In: Proc. 6th Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco

[17] Kemps-Snijders M (2010) RELISH: Rendering endangered languages lexicons interoperable through standards harmonisation. In: 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages, held in conjunction with LREC 2010, Valetta, Malta

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

[19] Leech G, Wilson A (1996) EAGLES recommendations for the morphosyntactic annotation of corpora. URL `http://www.ilc.cnr.it/EAGLES/annotate/annotate.html`, version of March 1996

[20] McCrae J, Spohr D, Cimiano P (2011) Linking lexical resources and ontologies on the semantic web with Lemon. The Semantic Web: Research and Applications pp 245–259

[21] McCrae J, Montiel-Ponsoda E, Cimiano P (2012) Integrating wordnet and wiktionary with lemon. Linked Data in Linguistics pp 25–34

[22] Pareja-Lora A, Aguado de Cea G (2010) Ontology-based interoperation of linguistic tools for an improved lemma annotation in Spanish. In: Proc. LREC 2010, Valetta, Malta

[23] Rehm G, Eckart R, Chiarcos C (2007) An OWL-and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In: Proc. RANLP 2007, Borovets, Bulgaria

[24] Santorini B (1990) Part-of-speech tagging guidelines for the Penn Treebank Project. Department of Computer and Information Science, University of Pennsylvania, technical report MS-CIS-90-47

[25] Saulwick A, Windhouwer M, Dimitriadis A, Goedemans R (2005) Distributed tasking in ontology mediated integration of typological databases for linguistic research. In: Proc. 17th Conf. on Advanced Information Systems Engineering (CAiSE'05), Porto

[26] Schiller A, Teufel S, Stöckert C, Thielen C (1999) Guidelines für das Tagging deutscher Textcorpora mit STTS. Tech. rep., Universities of Stuttgart and Tübingen

[27] Schmidt T, Chiarcos C, Lehmberg T, Rehm G, Witt A, Hinrichs E (eds) (2006) Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources, Proceedings of the E-MELD workshop on Digital Language Documentation

[28] Schuurman I, Windhouwer M (2011) Explicit semantics for enriched documents. What do ISOcat, RELcat and SCHEMAcat have to offer? In: Proc. 2nd Supporting Digital Humanities conference (SDH 2011), Copenhagen, Denmark

[29] Skut W, Brants T, Krenn B, Uszkoreit H (1998) A linguistically interpreted corpus of German newspaper text. In: Proc. ESSLLI Workshop on Recent Advances in Corpus Annotation, Saarbrücken, Germany

[30] Stede M (2004) The Potsdam Commentary Corpus. In: Proc. ACL-2004 Workshop on Discourse Annotation, Barcelona, Spain, pp 96–102

[31] Telljohann H, Hinrichs EW, Kübler S (2003) Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Tech. rep., Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany