Unity in Diversity: Integrating Differing Linguistic Data in TUSNELDA

Andreas Wagner

Universität Tübingen

This paper describes the creation and preparation of TUSNELDA, a collection of corpus data built for linguistic research. This collection contains a number of linguistically annotated corpora which differ in various aspects such as language, text sorts / data types, encoded annotation levels, and linguistic theories underlying the annotation. The paper focuses on this variation on the one hand and the way how these heterogeneous data are integrated into one resource on the other hand.

1 Introduction

The principal concern of the collaborative research centre (Sonderforschungsbereich) SFB 441 at University of Tübingen are the empiric data structures which feed into linguistic theory building. In order to approach this general issue from a considerable variety of research perspectives, SFB 441 comprises different projects each of which empirically investigates a particular linguistic phenomenon in a particular language or language family. The respective research interests range from suboptimal syntactic structures in German, local and temporal deictic expressions in Bosnian/Croatian/Serbian or Portuguese and Spanish, to semantic roles, case relations, and cross-clausal references in Tibetan, to mention just a few. As empirical basis for their research, many projects create electronically accessible collections of linguistic data and prepare them to fit their particular needs. In most cases, these collections are corpora. However, a couple of projects deal with data (e.g. lexical information) which are more adequately represented by an Entity-Relationship based data model and thus are implemented in relational databases rather than corpora.

Interdisciplinary Studies on Information Structure 02 (2005): 1–20

Dipper, S., M. Götze and M. Stede (eds.):

Heterogeneity in Focus: Creating and Using Linguistic Databases

All data collections built within SFB 441 projects are assembled in one repository called TUSNELDA (= *TUebinger Sammlung Nutzbarer Empirischer Linguistischer DAtenstrukturen*, Tübingen collection of reusable, empirical, linguistic data structures). Especially, the different corpora are integrated into a common XML-based environment of encoding, storage, and retrieval. This integration is particularly challenging due to the heterogeneity of the individual corpora, which differ with regard to the following aspects:

- languages (e.g. German, Russian, Portuguese, Tibetan,...)
- text types / data types (e.g. newspaper texts, diachronic texts, dialogues, treebanks, ...)
- categories of information covered by the annotation / annotation levels (e.g. layout, textual structure, morpho-syntax, syntax, ...)
- underlying linguistic theories

This paper describes the approach pursued to integrate these heterogeneous corpus data. Section 2 provides an overview of the corpora built by the individual projects. This overview illustrates the diversity of the data. Section 3 addresses their integration in TUSNELDA. In particular, aspects of the annotation process, the annotation schemes and the underlying data model, as well as corpus management and retrieval are discussed.

2 SFB 441 Corpora

This section provides an overview of the different corpora created in SFB 441. In the following listing, each project engaged in corpus building is mentioned together with the investigated language and the respective corpora. For each corpus, a short general description is given, including its size and a list of the annotation levels encoded in it.

Project A1: “Representation and automatic acquisition of linguistic data”
German

- **TüBa-D/Z (Tübinger Baumbank des Deutschen / Zeitungstexte)**
  manually annotated treebank (approx. 15,000 sentences)
  - syntactic structures

- **TüPP-D/Z (Tübinger Partiell Geparstes Korpus des Deutschen / Zeitungstexte**
  newspaper corpus; syntactically analysed by means of a rule-based chunk parser created in the project (approx. 200 million words; only partially integrated in TUSNELDA)
  - text structures (paragraphs, sentence boundaries, etc.)
  - syntactic structures

Project A3: “Suboptimal syntactic structures”
German

- **Database of Grammaticality Judgements**
  manually annotated example sentences originating from linguistic literature including grammaticality judgements (approx. 1,000 sentences)
  - morphological features (e.g. case)
  - syntactic structures

Project B1: “Corpus based study of address and linguistic politeness in the Slavonic languages”
Russian

- **Russian Interviews**
  interviews in newspapers (approx. 290,000 words)
  - text structures

- **Uppsala Corpus of Modern Russian**
  balanced Russian corpus compiled in Uppsala; extended by morpho-syntactic annotation by means of a POS tagger created in the project (1 million words)
  - text structures
  - morphological features / POS tags

Project B3: "Modal verbs and modality in German"
German

- **Goetz von Berlichingen**
  Early New High German text, digitised for the TITUS project (approx. 43,000 words)
  - text structure
  - layout (page and line breaks)

Project B8: "Corpus-based analysis of local and temporal deictics in (spontaneously) spoken and (reflected) written language"
Bosnian/Croatian/Serbian

- **Tübinger BKS-Korpus**
  Comic Corpus, Bosnian Interviews, Novosadski korpus of Spoken Language (approx. 127,000 words)
  - text structure / dialogue structure
  - marking and classification of deictic expressions
  - situational context (accompanying gesture)

Project B9: "Local and temporal deixis in the Romance languages — History and variation"
Portuguese, Spanish

- **TüPoDia (Tübinger Portugiesische Diachrone Texte)**
  Portuguese diachronic texts (approx. 260,000 words)
  - text structure
  - marking and classification of deictic expressions

- **BraToLi (Brasilien Toledo Lima)**
  transcriptions of spoken dialogs (including situational descriptions) from Brasil, Toledo, and Lima (approx. 10,000 words)
  - dialogue structure
  - marking and classification of deictic expressions
  - situational context

Project B11: "Semantic roles, case relations, and cross-clausal reference in Tibetan"
Tibetan

- **Tibetan Corpus**
  texts from different regions and epochs (currently approx. 700 clauses, to be extended)
  - text structure
  - layout (page breaks)
  - morphological features (e.g. case)
  - syntactic structures
  - verb–argument structures
  - cross-clausal references (anaphoric reference via empty arguments and pronouns)

For some of these corpora, substantial extensions are envisaged to cover additional annotation levels. For example, the German treebank TüBa-D/Z will be extended by co-reference annotation; the Tibetan corpus will be augmented by

lexical resources and English translations, which will be aligned to the annotated texts.

3 Integration in TUSNELDA

All the corpora mentioned in the previous section form the components (sub-corpora) of the TUSNELDA corpus. This means that they are integrated into a common environment regarding annotation, data management, and corpus querying. This environment is based on XML technology. This has two major advantages. Firstly, XML offers the flexibility required to encode all the peculiarities of the heterogeneous data sketched above. Secondly, various software for encoding, managing and querying XML documents is available and can be employed. The alternative, developing and implementing such software from scratch, appears infeasible in view of the diversity of requirements for encoding and processing the different corpora.

In detail, the integration of the different corpora involves several stages:

1. development of unified annotation schemes which cover all (combinations of) annotation levels realised in the TUSNELDA sub-corpora
2. transformation of the individual corpora into a format which obeys the respective annotation schemes
3. storing and managing the TUSNELDA sub-corpora in an XML database
4. implementation of query interfaces which are tailored to the respective annotation levels to be searched

3.1 Annotation Process

As noted in section 1, the individual sub-corpora of TUSNELDA are built separately in the respective SFB 441 projects. Moreover, their diversity implies

that different annotation procedures are most adequate and efficient in the respective corpus building activities. In this respect, two basic scenarios can be distinguished:

In one scenario, a proprietary data format and corresponding proprietary software is employed for annotation. This is appropriate in case there is an established way of annotating the information to be covered by the corpus, and in case a common and convenient annotation tool is available which supports this annotation. For example, the projects that create syntactic treebanks employ *annotate* (cf. Plaehn (1998)) for that task. This tool is widely used for building collections of syntactic trees. It provides a number of convenient features which speed up annotation, such as a graphical interface and facilities for interactive semi-automatic annotation. *annotate* encodes the data in the proprietary NEGRA format. A special case of this scenario is the use of tools for automatic annotation, such as POS taggers or shallow parsers, which of course require specific input and output formats. Integrating corpora built that way in TUSNELDA comprises two steps. Firstly, annotation schemes have to be developed and/or adapted to cover all information encoded in the corpora. Secondly, the corpora have to be converted from their respective proprietary format into XML structures which are conforming to the corresponding TUSNELDA annotation scheme. As a general rule, this format conversion can be done automatically.

In the alternative scenario, annotation immediately rests upon the TUSNELDA annotation schemes, i.e. TUSNELDA-conforming XML markup is created directly. This procedure is appropriate if a common practice for annotating the sort of information to be encoded in the corpus does not yet exist. Guided by their specific research interests, some projects create corpora which cover certain peculiar aspects (or combinations of aspects) for which neither an established annotation scheme nor a tailored annotation tool is available. For example, it was all but clear in advance how to adequately encode the closely interrelated aspects of syntactic structure, verb–argument structure and cross-

clausal reference in the Tibetan corpus. To handle such cases, a preliminary annotation scheme is developed in advance (as a DTD), and annotation is performed according to this scheme, using a general XML editor. In the course of the annotation process, with growing experience regarding the data, it usually turns out that revisions and extensions of the provisional scheme are necessary to appropriately encode certain peculiarities and/or to improve the possibilities of retrieving interesting information. Thus, the scheme is incrementally adapted to these emerging requirements. In this scenario, the annotation generally has to be performed manually. However, to increase efficiency, we aim at automatising annotation steps wherever possible (e.g. assigning unique IDs to elements). As annotation software we mainly use the CLaRK system (cf. Simov et al. (2001)), an XML editor which has been developed especially for encoding linguistic resources. On the one hand, this tool is not restricted to specific formats but supports any XML DTD. On the other hand, it comprises a number of facilities to perform annotation steps automatically or semi-automatically, such as regular grammar engines or constraint mechanisms which add specific markup depending on the context. These facilities are flexibly configurable and adaptable to the particular annotation scheme in use.[1]

3.2 TUSNELDA Annotation Scheme

Various general requirements guide the definition of annotation schemes for TUSNELDA. First of all, these schemes have to be exhaustive, i.e. they must capture all kinds of information which is encoded within the different annotation levels in the TUSNELDA corpora. As a second crucial requirement, the schemes should be convenient with respect to both annotation and retrieval. This means they should be designed in a way which facilitates manual anno-

[1] Wagner and Zeisler (2004) outline how these facilities are employed for annotating the Tibetan Corpus.

tation and allows the specification of “intuitive” search queries. These criteria imply two further requirements, which in a sense are complementary to each other. On the one hand, the schemes have to be open for different languages and linguistic theories. This is necessary since TUSNELDA is multilingual and its corpora are based upon differing theoretic approaches. On the other hand, analogous structures and phenomena in the different corpora should be encoded in analogous ways. This enhances reusability because it allows for the development of common mechanisms for annotation or format conversion as well as the implementation of analogous retrieval interfaces (including the specification of similar—if not identical—search queries) for the different corpora. In addition, keeping the annotation schemes as uniform as possible reveals commonalities and deviations of the information encoded in the different corpora.

Despite the diversity of the corpora in TUSNELDA, they all share the same generic data model: hierarchical structures. It is most appropriate to encode the phenomena captured in the TUSNELDA corpora by means of nested hierarchies, augmented by occasional “secondary relations” between arbitrary nodes in these hierarchies. This distinguishes TUSNELDA fundamentally from corpora whose annotation is based on other data models such as, for example, timeline-based markup of speech corpora or multimodal corpora (e.g. cf. Schmidt (2004)). Such corpora encode the exact temporal correspondence between events on parallel layers (e.g. the coincidence of events in speech and accompanying gesture or the overlap of utterances) whereas hierarchical aspects are secondary. In TUSNELDA, however, hierarchical information (e.g. textual or syntactic structures) is prevalent, while capturing the exact temporal coincidence of different events in general is not of primary relevance in the research within SFB 441.

Guided by these requirements, we decided to develop annotation schemes which encode information as embedded annotation (i.e. the markup is placed locally at or around the corresponding text) rather than standoff annotation (where

the markup is stored in a separate file, including pointers to the primary text). Essentially, this decision rests on two major considerations.

The first consideration concerns the required suitability of the schemes for manual annotation in particular and corpus processing in general. While standoff annotation appears to become a "quasi standard" paradigm for linguistic annotation, there is still a lack of general software supporting this paradigm. Usually, projects engaged in standoff annotation develop their own software which is tailored to their specific needs. Such software would, if at all, be only of limited use for annotating a corpus in TUSNELDA. Furthermore, due to the diversity of our corpora, we need general XML-aware tools which are adaptable to particular requirements of each individual corpus. Currently, such tools (XML editors, format conversion tools, XML databases and query engines) are optimised for processing hierarchical XML structures, i.e. they are well suited for embedded annotation, while providing at best rudimentary support for standoff annotation.

The second consideration is the fact that embedded annotation indeed is sufficient for encoding our data. Standoff annotation would be necessary if the structures to be encoded formed overlapping hierarchies, which cannot be modelled within a single XML document. Actually, this problem does not arise for our data. The structures primarily encoded in the TUSNELDA corpora are at the textual and/or syntactic level. Since syntactic structures constitute subsentential hierarchies while text structures define super-sentential hierarchies, these structures do not overlap so that they can be captured within a single document hierarchy. Concurrent hierarchical units occur only marginally and are not of primary importance. These units concern the physical (layout) structure of the annotated texts, e.g. page boundaries. Such boundaries are marked by empty XML elements (e.g. <pb/> for a page break), which do not violate the well-formedness of the document.

```
<s>
  <clause>
    <ntNode>
      <tok>
        <orth>khra·phru·gu</orth>
        <pos>NOM:anim∼pers</pos>
      </tok>
      <ntNodeCat>NP</ntNodeCat>
      <desc>
        <case>Abs</case>
      </desc>
    </ntNode>
    <tok id="v6">
      <orth n="2">med-tshug</orth>
      <pos>VFIN</pos>
      <desc>
        <feature type="part">NEG</feature>
        ...
      </desc>
    </tok>
    <clauseCat>simple</clauseCat>
  </clause>
  <punct>|</punct>
</s>
```

Figure 1: Example annotation from Tibetan corpus (1)

3.3 Examples

This section provides several examples which illustrate diverse (combinations of) annotation levels captured in the individual corpora and how these different sorts of information are encoded. These examples will also illustrate how the balance between the desired uniformity and the required flexibility w.r.t. different languages and theories is achieved.

Figure 1, taken from the Tibetan Corpus, exemplifies the encoding of syntactic structures. <tok> elements mark the tokens (i.e. words) of a text with their orthographic or phonemic realisation (<orth>) and part-of-speech classification (<pos>). A phrase is encoded by an <ntNode> (non-terminal node) element; <ntNodeCat> marks its category. For clausal constituents, there is a special element <clause> (including <clauseCat> specifying the clause category).[2] <ntNode> and <clause> elements may be recursively nested. Tokens, phrases, and clauses may receive a further linguistic description (<desc>). Such descriptions may contain simple features like case[3] or complex specifications like the argument structure of a verb.

An example for the encoding of argument structures in the Tibetan Corpus is shown in figure 2. This encoding belongs to the annotation displayed in figure 1. In fact it is located within the <desc> element of the verb token (at the position indicated by the dots) and presented here in a separate figure just for the sake of clarity. (This exemplifies the integration of different annotation levels—syntactic constituent structures and verb–argument structures—in one XML hierarchy.) In detail, the description comprises (a) the "canonical" argument structure (a list of <complement> elements within a <frame> element), and (b) the "real" frame, i.e. the realisation of the arguments in the clause, including additional arguments (a list of <realComplement> elements within a <realFrame> element). Each <complement> element within <frame> has a corresponding <realComplement> element within <realFrame> (possibly marked as not realised in the respective clause, see below). The order of <realComplement> items corresponds to the order of the respective <complement> items; additional complements which occur in the clause but are not included in the canon-

[2] In some corpora, no explicit distinction is made between clausal and other constituents; in these corpora, clauses are annotated as <ntNode> instead of <clause>.

[3] A certain set of common features is defined in the annotation scheme by specific elements such as <case>, <number>, or <person>. Furthermore, a general element <feature> (with a 'type' attribute) allows the specification of any kind of feature.

```
...
<frame>
   <complement>
      <role>POSS</role>
      <case>Aes</case>
   </complement>
   <complement>
      <role>EXST2</role>
      <case>Abs</case>
   </complement>
</frame>
<realFrame>
   <realComplement id="v6c1" status="empty">
      <role>POSS</role>
      <ref target="v5c1"> </ref>
   </realComplement>
   <realComplement id="v6c2">
      <role>EXST2</role>
   </realComplement>
</realFrame>
...
```

Figure 2: Example annotation from Tibetan corpus (2)

ical frame are represented by <realComplement> elements appended at the end of the <realFrame> list. In case the order of complements as realised in the clause deviates from the canonical complement order as defined in <frame>, <realFrame> receives the attribute 'order', which encodes the complement order in the clause (as a sequence of role labels).

For each canonical and real complement, the semantic role is specified. Furthermore, each canonical complement receives a specification of its case. The encoding of argument structure also captures information about cross-clausal references, especially the relation between empty arguments (i.e. arguments not

<figure id="*s45b3*">
  <figTrans>
    <sp who="*Komandant*">
      <spokenPar>
        Nadam se da govoriš istinu . . . Idite , potražite taoca , a <marked type="*deic-dem*">**ovu**</marked> **dvojicu u zatvor !**
      </spokenPar>
      <situation>
        <keywords>
          <term>*open hand*</term>
          <term>*stretched out*</term>
        </keywords>
      </situation>
    </sp>
  </figTrans>
</figure>

Figure 3: Example annotation from BKS Korpus (Comic Corpus)

overtly realised in a clause) and their antecedents in previous clauses.[4] To capture this kind of cross-clausal reference, each <realComplement> receives a unique ID. Empty arguments (e.g. the first <realComplement> in the example) receive an attribute marking emptiness and a pointer to the corresponding antecedent in the text, which in most cases is a <realComplement> specified in the argument structure of some previous clause. Such a pointer is encoded as a reference tag (<ref>) with an attribute 'target' that points to the ID number of the corresponding referee.

Figure 3 displays the encoding of a single comic picture in the BKS Comic Corpus. This encoding significantly differs from the previous examples in the

[4] The investigation of this phenomenon is one of the major research interests of project B11, which is building the Tibetan Corpus.

covered annotation levels; instead of entirely capturing complex syntactic structures, it provides punctual information about specific expressions (in this case deictics) and the situational context of their usage, especially accompanying gesture. A comic picture (captured by a <figure> element) is represented by a transcription (<figTrans>) of the dialogue taking place in this picture.[5] Each dialogue turn is encoded by a <sp> element with an attribute 'who' indicating the speaker. The utterance is captured by a <spokenPar> (spoken paragraph) element. Expressions of specific interest, as deictic expressions in the BKS Corpus, can be marked by the element <marked>; the attribute 'type' provides a classification of the expression. In the example, the word "ovu" is marked as demonstrative deictic ("*deic-dem*"). The element <situation> contains information about the situational context. In the Comic Corpus, this information is encoded as a set of keywords (a list of <term> elements within a <keywords> element) specifying gesture accompanying deictics. Note that this kind of transcription basically makes use of a hierarchical scheme rather than a timeline-based scheme employed for other transcriptions of dialogue. The research purpose which guided the creation of the Comic Corpus, i.e. the examination of deictic expressions and co-occurring pointing gesture, does not require the encoding of exact temporal overlaps between different utterances and/or nonverbal events. For this reason, the transcription of comics, where such temporal overlap is not determinable, is suitable for the research intended.

Figures 4 and 5 illustrate the openness of the TUSNELDA annotation schemes for different linguistic theories. Each of these figures shows a syntactic tree of a sentence: figure 4 from the TüBa-D/Z treebank, figure 5 from the Database of Grammaticality Judgements. Both sentences are in German and have considerable commonalities (wh-element "wie", "dass"-clause with

[5] More exactly, this transcription includes all written material, i.e. spoken utterances as well as text displayed on some artefact, e.g. a board, and "meta-situational" comments of the author located on top or bottom of the picture.

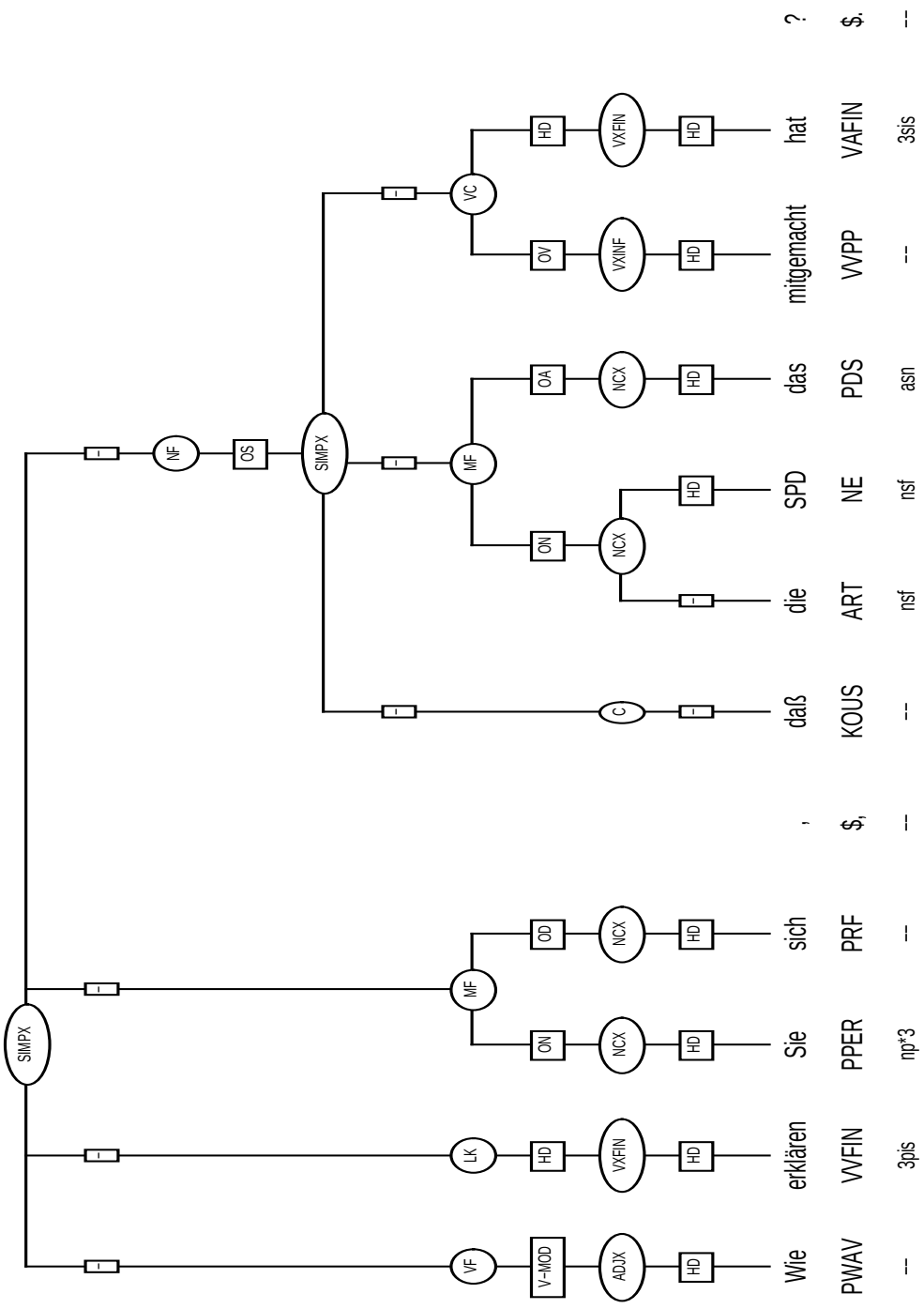

Figure 4: Example tree from TüBa-D/Z

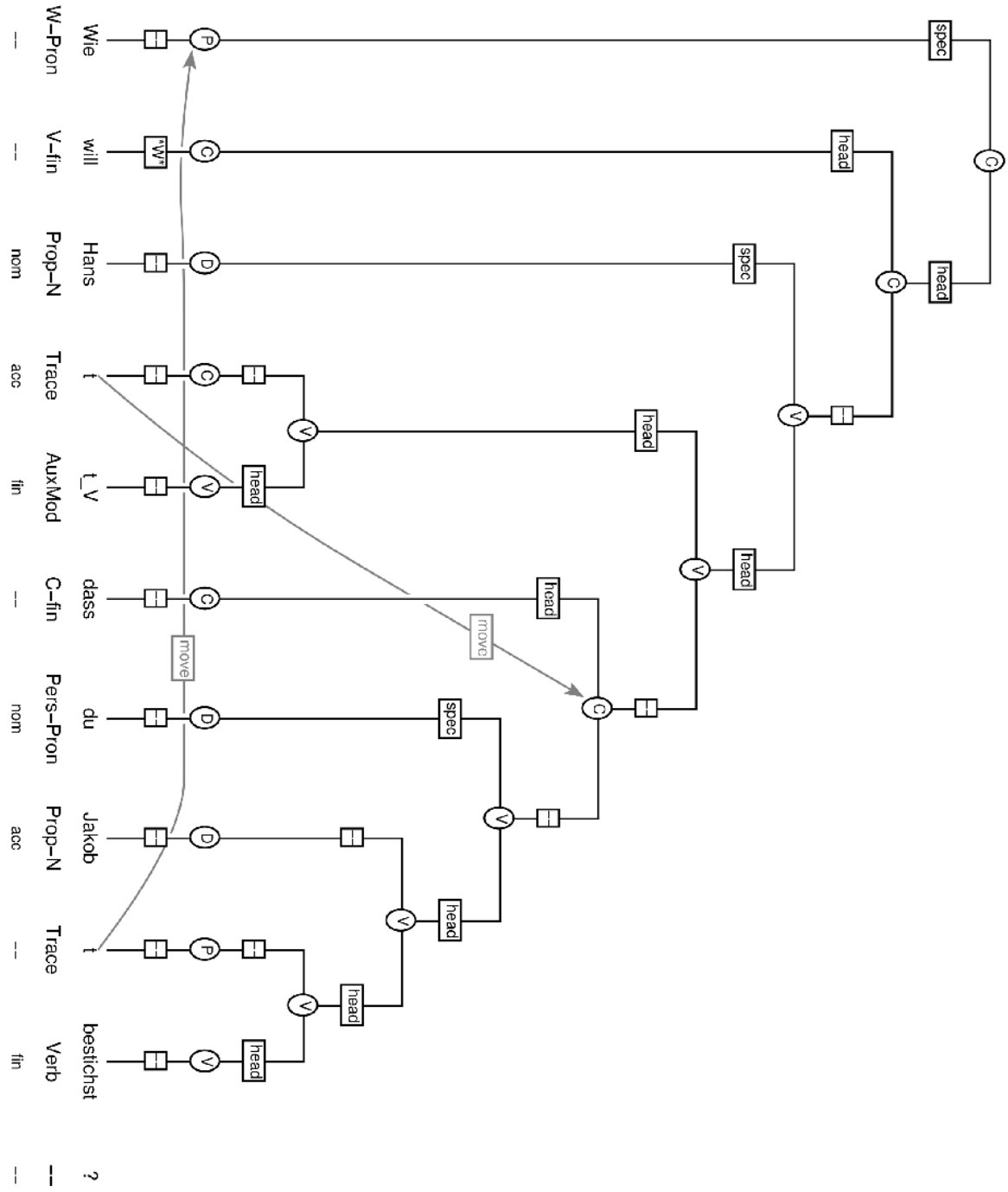

Figure 5: Example tree from the Database of Grammaticality Judgements

transitive verb). However, they are assigned very different syntactic structures, which reflect the linguistic theories and assumptions underlying the two treebanks. The annotation in TüBa-D/Z is guided by the theory of topological fields (a traditional descriptive theory accounting for the constituent order in German sentences) and the restriction to context-free structures, which results in comparably flat structures without traces. In contrast, the Database of Grammaticality Judgements is intended to comprise trees in accordance with generative syntax, characterised by highly nested (usually binary-branched) structures and the common use of traces. The TUSNELDA annotation scheme for syntactic structures is compatible to both approaches, i.e. both trees can be represented by an XML structure as in figure 1. The TUSNELDA scheme neither prescribes a set of POS tags and constituent labels nor constrains the configuration of syntactic trees. The only restrictions it imposes on the encoding of syntactic structures is the distinction between tokens (words) and non-terminal nodes (with the additional possibility to identify clause nodes by a special element) and the limitation to tree structures with possible secondary edges. These constraints mark the balance between the desirable uniformity and the required flexibility which is appropriate for TUSNELDA and its corpora.

3.4 Corpus Management and Querying

After the step of annotation (and, if necessary, format conversion), a corpus can be imported into a database which serves as the central platform for managing and querying the TUSNELDA corpora. As database software we employ *Tamino XML Server* developed by Software AG. Tamino is a native XML database and implements several techniques for indexing XML documents. This allows an efficient search in the data. Furthermore, Tamino provides a query language which is a subset of XQuery (cf. Boag et al. (in progress)). XQuery is being developed to serve as the standard language for querying XML data.

As Sasaki et al. (2004) point out, XQuery is particularly suited for retrieving hierarchical aspects of annotated material, which renders it less useful for corpora which are not based upon hierarchical data models. However, as discussed above, the annotation in TUSNELDA essentially is hierarchically organised so that XQuery is an appropriate query language.

The data in the TUSNELDA collection are made publicly accessible via a WWW interface (www.sfb441.uni-tuebingen.de/tusnelda.html). The Tamino software offers various facilities to configure HTTP-based interfaces for searching the XML database and formatting the query results. We employ these facilities to realise web interfaces which take into account the respective peculiarities of the individual corpora. The core of the search mechanism is the XQuery engine of the database. The user can formulate queries in a format based on XPath and XQuery. Concerning general accessibility of the interface, it makes more sense to rely on these standard languages for querying XML data than on proprietary query languages. However, the prospective users of TUSNELDA, i.e. linguistic and philological researchers, are usually not familiar with these languages. Therefore, we extend the interface with various mechanisms which render the interface more user-friendly. For instance, we provide corpus-specific example queries as well as templates and syntactic abbreviations which facilitate the formulation of "typical" queries. Furthermore, the user can choose between alternative formats of output display (e.g. syntactic structures can be viewed as graphical trees, labelled bracket structures, or XML structures as annotated in the corpus). Such facilities and their suitability to improve user-friendliness will be subject to the feedback by actual and prospective users inside and outside SFB 441. In this sense, the current WWW interface is in a preliminary state and will continually be refined to improve its benefit for the linguistic research community.

Bibliography

Scott Boag, Don Chamberlin, Mary Fernández, et al. XQuery 1.0: An XML Query Language. W3C working draft. Technical report, W3C, in progress. URL `http://www.w3.org/TR/xquery/`.

Oliver Plaehn. *Annotate Bedienungsanleitung*. Universität des Saarlandes, Sonderforschungsbereich 378, Projekt C3, Saarbrücken, Germany, April 1998.

Felix Sasaki, Andreas Witt, Dafydd Gibbon, and Thorsten Trippel. Concept-based queries: Combining and reusing linguistic corpus formats and query languages. In *Proc. of LREC 2004*, pages 259–262, Lisboa, May 2004.

Thomas Schmidt. Transcribing and annotating spoken language with EXMARaLDA. In *Proc. of LREC 2004 Workshop on XML-based Richly Annotated Corpora*, pages 69–74, Lisboa, May 2004.

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, and Atanas Kiryakov. CLaRK - an XML-based system for corpora development. In *Proc. of the Corpus Linguistics 2001 Conference*, pages 558–560, 2001.

Andreas Wagner and Bettina Zeisler. A syntactically annotated corpus of Tibetan. In *Proc. of LREC 2004*, pages 1141–1144, Lisboa, May 2004.

Andreas Wagner
Universität Tübingen
SFB 441
Nauklerstr. 35
72074 Tübingen
Germany
wagner@sfs.uni-tuebingen.de
http://www.sfb441.uni-tuebingen.de/~wagner