Querying both time-aligned and hierarchical corpora with NXT Search

Ulrich Heid*, Holger Voormann*, Jan-Torsten Milde$, Ulrike Gut°, Katrin Erk†, Sebastian Padó†

*Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany
{heid, voormann}@ims.uni-stuttgart.de

$Fachhochschule Fulda, Marquardstrae 35, 36039 Fulda
Jan-Torsten.Milde@fh-fulda.de

°Englisches Seminar, Albert-Ludwigs-Universität Freiburg i.Br., Rempartstr. 15, 79098 Freiburg, Germany
ulrike.gut@anglistik.uni-freiburg.de

†Computerlinguistik, Universität des Saarlandes, Im Stadtwald 17, 66123 Saarbrücken, Germany
{erk, pado}@coli.uni-sb.de

Abstract

One problem of the (re-)usability and exchange of annotated corpora is in the lack of standards in corpus formats and corpus query tools. This paper reports on the NXT Search tool, which was used to query two corpora with very different annotation formats. It is shown that with automatic data format conversion both corpora can be accessed and searched with NXT Search.

1. Introduction

An increasing number of corpora is being produced worldwide which are annotated at several levels of linguistic description. Such multi-level corpora may consist of, e.g., speech recordings with phonetic and prosodic annotations and annotations of the word class of each word form, but also several types of text-based annotations, e.g. covering, e.g., (morpho-)syntax, lexical semantics, or coreference. Examples of such multi-level corpora include MapTask[1], the LeaP (cf. (Milde/Gut 2002a)) and SALSA (cf. (Erk et al. 2003)) corpora dealt with in this paper, as well as most multimodal corpora.

One of the research benefits associated with multi-level corpora is the fact that, in principle, they allow linguists to observe the cooccurrence of phenomena from different levels of linguistic description, within a large enough context. This is necessary to explore the data and to check them for consistency. However, this kind of corpus use is not easy to realize, as there are few general query tools, and sharing or exchange of corpora almost inevitably means also the sharing or exchange of proprietary tools.

In this paper, we claim that adherence to a small number of basic guidelines and the use of a sufficiently general query tool greatly support access to cross-level linguistic data, on the basis of an exchange or shared use of multi-level corpora. We support this claim by reporting on two experiments: two multi-level corpora designed according to two quite diverging research traditions were both converted into the format used by the NXT Search query tool (Voormann et al. 2003). One of them, the LeaP corpus, follows the corpus design principles in use in speech research and conversational analysis (flat, time-aligned corpora), the other, the SALSA corpus, is based on a treebank format and contains multiple intersecting hierarchical annotations.

Section 2 contains an overview of corpus models and tools, and of our source corpora. In section 3, we describe their conversion, in section 4 a few sample queries. The guidelines for corpus design are discussed in section 5.

2. Corpus Models and Corpus Query in Speech- and Text-based Research

2.1. Corpus Models

Most speech corpora are time-aligned; they contain events annotated with start and end time stamps anchored to the recording. They typically have a flat structure, i.e. they have non-hierarchical point-wise or interval-related annotations. Examples of annotation tools producing such formats include ESPS/waves+, Praat, Anvil or the TASX annotator (Milde/Gut 2002b).

In text-based natural language processing, corpora typically are not time-aligned, but rather use word forms as a basic unit to which annotations are anchored. Besides of single words, also word groups, chunks and recursive, hierarchical structures are annotated, as e.g. in tree banks. Example formats include the University of Pennsylvania Treebank[2], or the German TIGER treebank (Mengel/Lezius 2000)[3].

2.2. The Corpora used and their Tools

2.2.1. The SALSA Corpus

The SALSA corpus ("The *SA*arbrücken *L*exical *S*emantics Annotation and *A*nalysis Project"), is currently based on about 1.5 million words of German newspaper text (*Frankfurter Rundschau*, 1992/93). It extends parts of the TIGER treebank (annotated with words, lemmas, parts of speech, grammatical categories and grammatical func-

[1] www.ldc.upenn.edu/Catalog/readme_files/hcrc.readme.html

[2] http://www.cis.upenn.edu/ treebank/home.html

[3] http://www.ims.uni-stuttgart.de/projekte/TIGER/

tions) with semanic frames and roles in the style of Frame Semantics, as used in the FrameNet project.

Frames in Frame Semantics are representations of prototypical situations, introduced by lexical predicates (verbs, nouns or adjectives). Frame instances are represented as trees of depth one, with node labels indicating frame names (e.g. *categorization* in figure 1) and edge labels indicating role names (e.g. *Item* and *Category* in figure 1). The edges of a frame tree point to nodes of the syntactic structure which results in an annotation with intersecting syntactic and semantic trees.

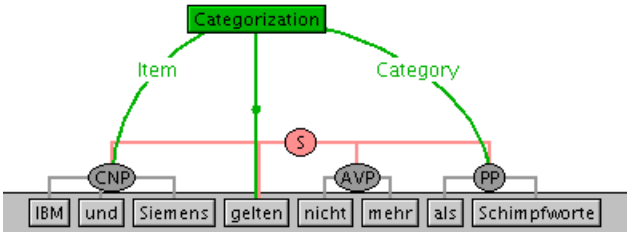

Figure 1: SALSA browser: annotated corpus sentence

The SALSA corpus is represented in the TIGER/SALSA XML format, a modular extension of the TIGER XML format (Mengel/Lezius 2000). Semantic and syntactic annotations are stored separately from each other, with reference through unique IDs. The format supports reference beyond sentence boundaries, as well as below word boundaries (i.e. into parts of, e.g., compounds); it also supports semantic underspecification.

SALSA data can be viewed with a specific corpus browser. So far, however, there is no general query tool available which allows easy access to cross-level observation data.

2.2.2. The LeaP Corpus and the TASX Format

The LeaP corpus ("*Lea*rning *P*rosody") was collected in the *Learning Prosody* project (cf. (Milde/Gut 2002a)), from native as well as non-native speakers of German and English. The data (a total of over 12 hours of speech, 359 recordings) consist, among others, of readings and retellings of a short story, as well as of interviews.

The LeaP data are annotated at eight different linguistic levels comprising speech vs. non-speech, words, syllables, segments, intonation, pitch movements as well as part of speech and lemma values of word forms (cf. figure 2).

The LeaP corpus is represented in the TASX format (Milde/Gut 2002b), a flat XML-based corpus representation format. TASX corpora may consist of an arbitrary number of sessions (equivalent, in LeaP, to recordings), each of which may hold several descriptive tiers. Each tier in turn consists of a separate set of events (in LeaP, each of the eight descriptive levels is encoded in a tier). Events have some textual information and are linked to primary data (e.g. audio files) via start and end time stamps. TASX data may in addition carry metadata at session, tier or event level.

Tools for the work with corpora in the TASX format include the TASX visual annotator for multimodal cor-

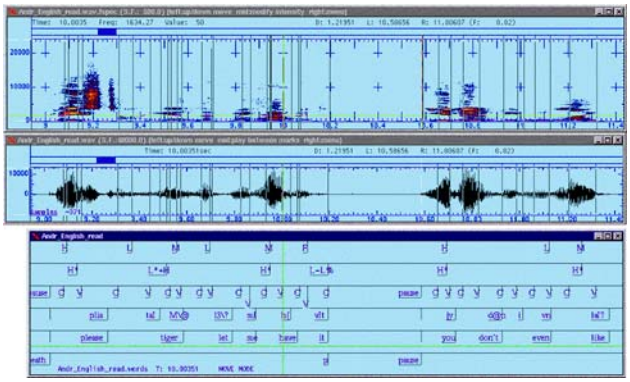

Figure 2: The six manual annotation tiers in LeaP.

pora, an interactive input panel for gesture types, a corpus browser (with string search applicable to one tier at a time), as well as specific extraction tools based on Perl and/or XSLT. Recently, an XQuery-based library of (often used) queries has been constructed, but there is no general query tool so far supporting the provision of cross-level observation data in a way easily accessible to linguists.

2.3. Query Tools

For both corpus traditions there are specific query tools. Many speech corpora can either be queried with inbuilt scripts (e.g. Clan analysis commands[4] or the EMU system [Cassidy & Harrington 2001]) or with custom-made scripts for statistical postprocessing. Similarly, there are query tools for treebanks, such as *tgrep* for the U Penn Treebank or *TIGERSearch* (Lezius 2002) for TIGER. Both allow to query single hierarchies, but not overlapping ones.

As most multi-level corpora are represented in XML, XML-aware query can be applied to them, based on the *XPath* or *XQuery* standards. An alternative are XML databases, such as TAMINO[5] or Xcerpt[6]. Even though these tools provide full flexibility for searching data, they are probably not very intuitive for linguists. Most queries in, e.g. XQuery or Xcerpt, are written in the form of programs with control structures and value assignments.

The recent Annotation Graphs approach and the pertaining query language (Bird et al. 2002), even though promising in principle, may also be suboptimal for the use by linguists, as it does not preserve either of the two corpus models described above in section 2.1.: instead, it maps any incoming corpus model onto a graph model, where, for example, corpus partitions such as sessions, turns, sentences, etc. are not immediately accessible.

3. Converting LeaP and SALSA to NXT

In the conversion of the LeaP and SALSA corpora, a major objective was to carry out as much as possible of the conversion automatically. Even though in principle any standoff architecture would allow this convertion, we are not aware of any published work of this kind.

[4] http://childes.psy.cmu.edu/clan/
[5] http://www.softwareag.com/tamino/
[6] http://www.xcerpt.org/

3.1. The NXT Corpus Format

The NXT data model (developed as part of the NITE XML Toolkit, within the EU project NITE[7]) is based on the principles of standoff corpora and combines sequential and hierarchical structures, allowing the representation of both, time-aligned and hierarchically annotated data. It supports intersecting hierarchies: a given element may have as parents elements from different hierarchies. The NXT object model (NOM) is serialized into a collection of mutually linked XML files.

Elements of NXT may have arbitrary text attached, as well as start and end time stamps. Furthermore, elements can inherit time information from their children.

3.2. Converting SALSA to NXT

The SALSA corpus format and the NXT corpus format are similar insofar as the XML hierarchy is extended by additional parent-child relations. The transformation between SALSA and NITE is fully implemented in XSLT and creates three XML files: part of speech, syntax and semantics. The corpus file, which describes the structure of the generated NITE corpus, was written once and fits all NXT SALSA (sub)corpora produced by this XSLT script.

3.3. Converting LeaP to NXT

Equally, the LeaP to NITE conversion is fully implemented in XSLT. In the first step, the TASX-annotated LeaP corpus is split up into eight files, each storing a single tier, creating the stand-off markup proposed by NITE. In addition, the conversion tool generates an XML-annotated corpus file. The conversion process is further controlled by a configuration file, which describes the hierarchical relations of TASX-annotated layers. Hierarchical relations between layers are computed on the basis of temporal overlap of their elements. If elements of one layer (e.g. the syllable layer) are fully contained in the temporal interval of an element on another layer (e.g. the word layer), it is assumed that a hierarchical relation between these elements can be constructed. Consequently, the element on the word layer will be the parent of the elements on the syllable layer. However, it has to be taken into account that manual annotations on different tiers using software such as ESPS/waves+ never result in perfectly matching time stamps even when the visualisation suggests perfect alignment. In the configuration file, a limit of tolerance of an overlap can be specified.

4. Querying LeaP and SALSA with NXT Search

4.1. NXT Search

The NXT Search query tool (NXT Search) is based on the NXT data model. It is implemented in Java, with a prototype available for free.

The NXT Search query language provides attribute tests (including regular expressions), and queries for temporal and structural relations. Attribute tests concern values of annotations from any linguistic level, as well as start and end time points, and the duration of intervals. Temporal relations may be used to compare the temporal extension of two events, with operators for precedence, temporal overlap, inclusion, identity, and start or end alignment. Structural relations include dominance (in a hierarchical annotation) as well as precedence (two elements are in a precedence relation, if they have a common ancestor element).

NXT Search queries consist of a variable declaration (first line in the example below) and a (regular expression of) matching condition(s). Variables are marked by the "$" sign and declared with respect to the level of annotation to which they belong (i.e. to the respective element type). The sample query in figure 3 searches for words *$w* which dominate syllables *$s* that contain a schwa (annotated as "@" in the *value* attribute).

4.2. Querying LeaP and SALSA

In the conversion, not only no information encoded in the original LeaP and SALSA corpora was lost, but also the hierarchical structure of some of the LeaP annotation tiers (e.g. *words* include *syllables*) was made explicit, and thus queriable.

We now show sample queries for both corpora. We search for cooccuring phenomena from different levels of description. An export of the hits for subsequent statistical treatment is possible as well.

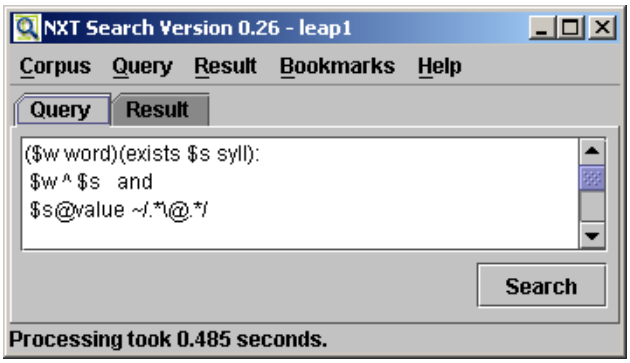

Figure 3: NXT Search

Within LeaP we searched for pitch accents on non-content words (conjunctions, determiners, prepositions or the word *to*) in non phrase-final position. This query combines information from speech- and text-based levels. We expected only few results in native speech (except for contrastive accent), but more in non-native speech. Table 1 presents the results for the retelling by a non-native speaker. He produces more non-content words with pitch accents than the native, especially conjunctions (cf. (Milde/Gut 2002a))

A typical cross-level query in the SALSA corpus is the following: find words or syntactic categories (e.g. NP, PP,...) which are the target of different semantic frames or which have more than one role, each role belonging to a different frame. An example of such a case is depicted in 4: the verb *kommen* ("arrive") is the target (i.e. frame-evoking element) of two frames: "arriving" and "operate-vehicle". Similarly, the subject NP of the sentence, *"die modernen*

[7] http://www.ims.uni-stuttgart.de/projekte/nite/

word	pos	context
and	conjunction	a piece of cheese **and** he want
because	conjunction	can not eat it and **because** the tiger
but	conjunction	there was a a frog **but** he also saw no chance
that	conjunction	make the cheese smaller **that** the tiger can eat it
at	preposition	nibble **at** the ch cheese
that	conjunction	... and **that** the tiger can eat the cheese

Table 1: Pitch accents on non-content words: results non-native speaker

Goldgräber" fills two roles: the *theme*-role of the *arriving*-frame and the *driver*-role in *"Operate_vehicle"*. The respective NXT Search query is formulated as follows:

```
($f1 frame)($f2 frame)
(exists $phrase syntax)
(exists $target word):
  $f1 >"target" $target and
  $f2 >"target" $target and
  $f1 ^ $phrase and
  $f2 ^ $phrase and
  $f1 != $f2
```

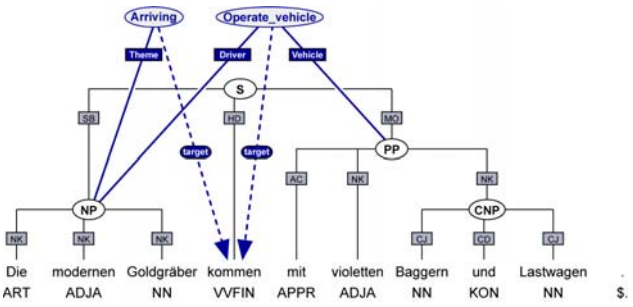

Figure 4: SALSA: Sample result of the query

5. Conclusions

Our experiments have shown that it is possible to convert both time-aligned corpora and corpora with intersecting hierarchical annotations to the NITE NXT Corpus format. The conversion can be carried out fully automatically (by means of an XSLT stylesheet) for SALSA and mostly automatically with the time-aligned LeaP corpus. In the latter case, the making explicit of implicit hierarchical structures requires manual intervention in the creation of a configuration file that guides the (otherwise automatic) conversion from the TASX format.

The NITE NXT data model serves as a metamodel for both corpora NXT Search is a query tool adapted to the NITE data model and allows a flexible querying of these types of corpora.

One prerequisite for a successful conversion is the adherence of the source corpora to a few general principles for XML-based corpus modelling:

- Annotations from different levels of linguistic description should be kept separate.
- Where appropriate, events should bear explicit time stamps for start and end time of the event.
- Hierarchical relationships between elements of the annotations should be made explicit.

If these principles are observed in the definition of corpus formats, a conversion should be possible also for multimodal corpora.

NXT Search is a laboratory prototype. Even though it is quite flexible, a few operators are still lacking (fuzziness at time relations and distance specification at structural precedence relations). For use by linguists, more (linguistically adequate) abstraction in the query language would be useful. NXT Search, as well as all other tools we examined, are not yet optimized for handling large amounts of data: for example, the entire LeaP corpus cannot be loaded at once by any of them.

Future work should address these issues: more flexibility and more abstraction in query languages for multi-level corpora, and optimized query tools. In addition, corpus construction should be based on the guidelines mentioned above.

6. References

Bird, S., P. Buneman, W.-C. Tan, 2002. Towards a query language for annotation graphs. *Proceedings of the Second International Conference on Language Resources and Evaluation*, Las Palmas, 807-814

Cassidy, S., and J. Harrington, 2001. Multi-level annotation in the Emu speech database management system in: *Speech Communication*, 33:61–77.

Erk, K., A. Kowalski, S. Padó, and M. Pinkal, 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. *Proceedings of ACL-2003*, Sapporo.

Lezius, W., 2002: *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis IMS, University of Stuttgart; AIMS, volume 8, number 4.

Mengel, A., and W. Lezius, 2000. An XML-based encoding format for syntactically annotated corpora. *Proceedings of LREC-2000*, Athens.

Milde, J.-T., U. Gut, 2002a. A Prosodic Corpus of Non-Native Speech. B. Bel and I. Marlien (eds.), *Proceedings of the Speech Prosody conference*. Aix-en-Provence: LPL, 503–506, http://leap.lili.uni-bielefeld.de

Milde, J.-T., U. Gut, 2002b. The TASX environment: an XML-based toolset for time aligned speech corpora. *Proceedings of the third International Conference on Language Resources and Evaluation*, Las Palmas, 1922–1927.

Voormann, H. et al. 2003. *NXT Search User's Manual (Draft)*, http://www.ims.uni-stuttgart.de/projekte/nite/manual/