Querying Annotated Speech Corpora

Ulrike Gut, Jan-Torsten Milde†, Holger Voormann° & Ulrich Heid°*

*University of Freiburg, Germany
†Polytechnical University Aalen, Germany
°IMS, University of Stuttgart, Germany
ulrike.gut@anglistik.uni-freiburg.de, Jan-Torsten.Milde@fh-aalen.de,
{holger.voormann;ulrich.heid}@ims.uni-stuttgart.de

Abstract

This paper is concerned with querying annotated speech corpora. A growing number of such corpora is currently being created worldwide; however, their usefulness for a wider research community is restricted by the lack of standard tools for creating, editing, annotating, storing and querying them. Two solutions for these problems are presented here: the XML-based data format TASX for corpus creation and data format exchange and the NXT search tool for querying corpora. Both tools have been applied to the multi-level annotated LeaP corpus of non-native speech.

1. Introduction

A growing number of annotated speech corpora is being produced all over the world following the demand for both technological applications and the development of theoretical models of spoken language. Whereas commercial applications of *prosodically* annotated corpora are still rare [18,25], linguistic research is increasingly making use of them. This includes branches of linguistics such as prosodic typology [14], sociolinguistic variation [7,10], prosody and discourse structure [23] and second language acquisition [8,17]. The widely recognized potential of this methodology includes the heuristic power of corpus searches that support the generation of new hypotheses about language and reveal previously unsuspected linguistic phenomena.

It is, however, still characteristic of corpus-based research that in the different projects use is made of a wide range of different tools for corpus creation, annotation, storage and query, very often specifically developed for only one particular project. In order to make the corpora available to a large number of users outside the respective project and in order to open up the corpus for queries not originally intended by the corpus creators, standard tools and corpus formats need to be developed for the following aspects [see also 1]:

- **Annotation tools:** a variety of annotation software such as ESPS/waves+, Praat, Anvil [13] and the TASX-annotator [16] exist, which support multi-level prosodic annotation of speech and even video data. Most of them are based on a proprietary data format which is not easily transformable into other data formats.
- **Corpus data format:** A proprietary data format precludes the use of the corpus by other researchers. A standardised data format is needed to allow the exchange of data.
- **Query tools:** A variety of query tools and query languages have been developed (e.g. [3]), but are usually limited to the respective corpus data format.

In this paper, we will describe a method of how annotated speech corpora can be made accessible for a wide range of research purposes. We will present the XML-based data exchange format TASX that serves as a conversion format for a great number of data formats used in the various extant speech corpora. In addition, this data format can be converted into the NITE data format so that the NXT search tool can be used to query the corpora.

The paper is structured in the following way: In Section 2, the multi-level annotated LeaP corpus of non-native speech will be described in terms of data and annotation and the XML-based data format TASX is presented. The NXT search tool and the NXT query language are described in Section 3, and Section 4 illustrates how the corpus can be queried.

2. The LeaP corpus

The LeaP corpus was collected in the LeaP (Learning Prosody) project (cf. <http://leap.lili.uni-bielefeld.de>), which was concerned with the acquisition of prosody by non-native speakers of German and English. The aims of the project include both the phonetic and phonological description of non-native prosody and the exploration of learner variables that influence the acquisition process. Data was collected from different groups of speakers: learners before and after a period abroad, before and after a four-month prosody training course, especially advanced learners who are hardly distinguishable from native speakers, and learners with different levels of competence. A quasi-experimental study was carried out which compared a treatment group of students taking part in a theoretical and practical training course in prosody with a control group.

In addition, a large number of meta data was collected for each recording including meta data about the recording: (date, place, interviewer and language of the interview), meta data about the non-native speaker (age, sex, native language/s, second language/s, age at first contact with target language, type of contact, (formal vs. natural), duration and type of stays abroad, duration and type of formal lessons in prosody (if at all), prosodic knowledge) and meta data about motivation and attitudes (reasons for acquiring the language, motivation to integrate in the target country, attributed importance to competence in pronunciation compared to other aspects of language, interest, experience and ability in music and in acting).

2.1. Data

The recordings consist of readings of nonsense word lists and three different speech styles:

- Readings of a short story (about 2 minutes).
- Retellings of the same story (between 2 and 5 minutes).
- Interviews (between 10 and 30 minutes).

The entire corpus consists of 359 annotated files and includes a total of 131 different speakers with 32 different native languages as well as 18 recordings with native speakers. The total amount of recording time is more than 12 hours.

2.2. Annotation

The annotation comprises phonological and prosodic as well as morphosyntactic events on eight different tiers:

- On the phrase tier, speech and non-speech intervals are transcribed. Non-speech events include unfilled pauses, noise, breath, laughter and hesitation phenomena. Speech events are transcribed as complete intonational phrases or as interrupted phrases. Elongated phonemes are marked as well.
- On the syllables tier, syllables are transcribed in SAMPA [24]. The determination of syllable boundaries is based on auditory criteria which allow for resyllabification processes in spoken language [6].
- On the segments tier, all vocalic and consonantal intervals plus the intervening pauses are annotated; vowels and postvocalic semi-vowels are considered vowels and plosives, fricatives, nasals, approximants, affricates, prevocalic semivowels, laterals, trills and retroflexes are considered consonants. The determination of the vowel boundaries is supported by a broad band spectrogram and carried out following phonetic standard criteria (e. g. [19]).
- On the tones tier, pitch accents and boundary tones are annotated using a modified version of ToBI [9,22].
- On the pitch tier, the initial high pitch, the final low pitch and intervening high peaks and low valleys are annotated.
- On the words tier, words are transcribed orthographically. Marking of cliticizations such as "aren't" is possible.
- On the POS tier, the parts-of-speech are annotated automatically. For German, the IMS decision tree tagger [21] and for English the Penn Treebank is used [15].
- Lemmatization is carried out automatically during the tagging process.

For a recording of about one minute length, on average, 3000 events are annotated.

2.3. Data format

All data in the LeaP corpus is stored in an XML-based format called TASX: the Time Aligned Signal data eXchange format. It is increasingly accepted that current standard XML technology (i.e. XML, XSLT, XPATH, XQUERY) can be used to model linguistic corpora, to transform, query and distribute the content of such corpora and to perform adequate linguistic analysis [1]. Using XML as a data format, language data is stored in the form of tree-structured text files. A separate, formally defined document grammar can be used in order to test the structural correctness of a document in a validating XML-parser. Optionally, the document can be converted into an internal representation by the parser. There are further approaches for the processing of XML-structured data, which are partly standardised by the W3C or ISO.

A TASX-annotated corpus consists of a set of sessions, each one holding an arbitrary number of descriptive tiers, called layers. Each layer consists of a set of separate events. Each event stores some textual information (e.g. a syllable) and is linked to the primary audio data by two time stamps [16,17]. Arbitrary meta-data can be assigned to the complete corpus, each session, each layer and each event. In the LeaP corpus, each recording is encoded as a session with one layer for each of the eight annotation tiers. The beginning and end of each event (e.g. a word) is referred to in the respective time stamps. Events that occur only at a point of time instead of an interval (e.g. a pitch accent or the final low of the pitch curve), are referred to by only one time stamp. The meta data is encoded in an extended IMDI/ISLE format and integrated into the TASX file.

Despite its simplicity, the TASX-format is powerful enough to encode most of the corpus annotation formats currently in use. Indeed, a number of format transformation programs have been implemented, for example for annotation files produced by Praat, ESPS/waves+, SyncWriter, Exmeralda and Anvil.

The TASX format meets the requirements for a general exchange format for multi-level multi-modal language data. However, it is not optimised for storing or querying corpus data. This becomes evident when queries across tiers have to be expressed. While it is always possible to use XSLT/XPath [12] to formulate the query operations, the syntactical complexity of such queries is too high for standard users such as linguists. Equally, the simplified query syntax proposed in [4], which is similar to XQuery [11], is still complicated to learn and apply. The NXT search tool provides an abstract query language with which queries can be made in a precise fashion.

3. The NXT Search tool

The NXT search tool (NXT Search) is a component of the NITE XML Toolkit (NXT) and was developed in the NITE (Natural Interactivity Tools Engineering) project (cf. <http://nite.nis.sdu.dk/>). NXT includes Java libraries and the specification of the underlying XML-based data model for annotating, editing, visualising and searching multi-level, cross-level and cross-modality data. The data model supports intersecting hierarchies; for example, a hierarchical semantic structure can be bound to elements of a syntax tree in such a way that an element may have both a semantic and a syntax element as parents [5]. This acyclic graph is serialized into a couple of mutually linked XML files. Like in TASX, in NXT elements may carry time stamps. Furthermore, elements can inherit time information from their children.

The query language of NXT Search provides attribute tests (including regular expressions), structural, and temporal relations. For example, one can search for words *$w* containing a syllable *$s* (elements of the word tier dominating ^ an element on the syllable tier), which contains a schwa (annotated as '@' character in the value attribute):

$$(\$w\ word)(exists\ \$s\ syll):\quad \$w\text{^}\$s\ \ and\ \ \$s@value \sim /.*\backslash@.*/ \tag{1}$$

3.1. TASX-NITE data conversion

In order to be able to use the NXT search tool for the LeaP corpus, a conversion process between TASX and NITE has been implemented (*TASX2NITE.xsl*), which automatically transforms TASX-annotated language data into the XML-annotated NITE file format. The converter is fully implemented in XSLT so that it is also possible to use the converter directly from within the TASX-annotator and to export sessions directly to NITE.

In the first step of the conversion, the TASX-annotated LeaP corpus is split up into eight files, each storing a single annotation layer, which references the primary data via *nite:start* and *nite:end* attributes. By this the stand-off markup proposed by MATE/NITE [2] is created. In addition, the conversion tool generates an XML-annotated corpus file, which describes the structure of the generated NITE corpus (see Figure 1).

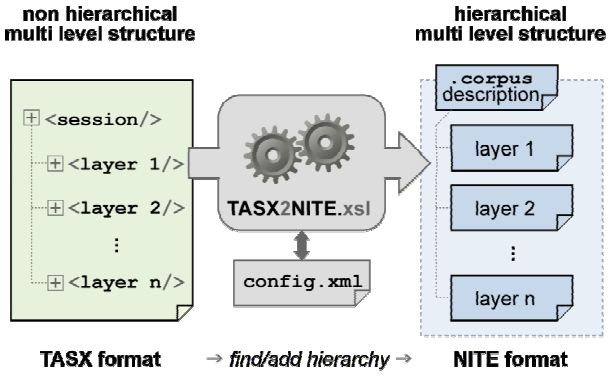

Figure 1: *The conversion process.*

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. For those elements of one layer (e.g. the syllable layer) that 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 with the variable *epsilon*. For example, a tolerance interval of 100 ms was specified for the end of a syllable and the end of the corresponding word label.

A major advantage of this approach is that it now becomes possible to use the time aligned, non hierarchically annotated data common in the linguistic fields of phonetics or conversational analysis and to generate hierarchical structures based on linguistic content and/or time relations and to finally apply hierarchy based search and transform operations, which, up to now, were mainly used in linguistic fields like syntax or semantics. The following section shows examples of such combined queries.

4. Querying the LeaP corpus

A variety of different types of queries of the LeaP corpus have been carried out with the NXT search tool. A simple query, for example, combines information from two different annotation tiers, tones and words. In Figure 2, a search for all nouns (*pos="NN"*) at the end of an intermediate phrase ending in a low tone (*tones="L-"*) is exemplified. Tones refer to points in time (provisionally encoded as end points of trivial intervals); temporal inclusion is expressed by the statement in line 3 of Figure 2.

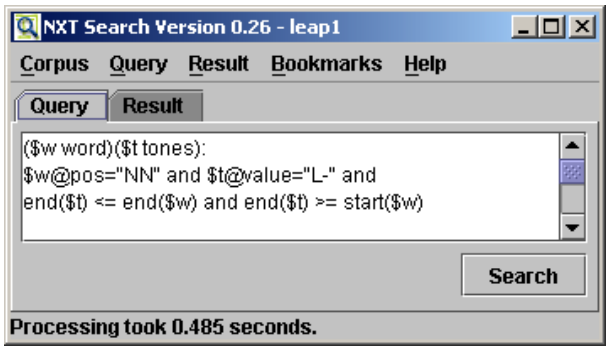

Figure 2: *Query for nouns at the end of an intermediate phrase ending in a low tone.*

With another query we search for pitch accents on non-content words (conjunctions, determiners, prepositions or the word *to*) in non phrase-final position:

$$
\begin{aligned}
&(\$w\ word)(exists\ \$wNext\ word)(exists\ \$t\ tones):\\
&\quad(\ \$w@\,pos="CC"\ \ or\ \ \$w@\,pos="DT"\ or\\
&\qquad\$w@\,pos="IN"\ \ or\ \ \$w@\,pos="TO"\ )\ and\\
&\$t@value \sim /.*\backslash*/\ \ and\\
&end(\$t) <= end(\$w)\ \ and\ \ end(\$t) >= start(\$w)\ \ and\\
&\$w\ ][\ \$wNext\ \ and\ \ not\ \$wNext@value=""
\end{aligned}
\qquad (2)
$$

We expected those cases to be extremely rare in native speech but more frequent in non-native speech. In one of the native speaker retellings we find four hits (see Table 1). However, as can be seen in d), the pitch accent on *this* has the function of a contrastive accent.

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

	word	pos	context
a)	**or**	conjunction	no, hop it [pause] **or** I'll eat you too
b)	**on**	preposition	and went **on** his way
c)	**into**	preposition	offered to help by jumping **into** his mouth
d)	**this**	determiner	and at **this** moment, …

Table 2 presents the results for the retelling by a non-native speaker. First, it can be seen that he produces more non-content words with pitch accents compared to the native speaker. Second, other word categories are affected: in nearly all cases conjunctions have pitch accents, whereas for the native speaker prepositions were predominately affected.

It is also possible to use NXT Search for the evaluation of the syntactic correctness of the corpus annotation. For example, the annotation schema of the LeaP project requires that intonation phrases be separated by pauses or other non-speech events. In a trial search for other combinations, one erroneous sequence of three intonation phrases was identified.

Table 2: *Pitch accents on non-content words: results non-native speaker.*

	word	pos	context
a)	**and**	conjunction	a piece of cheese **and** he want
b)	**and**	conjunction	to eat the cheese **and** the cheese was …
c)	**be-cause**	conjunction	can not eat it and **because** the tiger
d)	**but**	conjunction	there was a a frog **but** he also saw no chance
e)	**that**	conjunction	make the cheese smaller **that** the tiger can eat it
f)	**at**	preposition	nibble **at** the ch … cheese
g)	**that**	conjunction	… and **that** the tiger can eat the cheese

5. Outlook

This study has shown that XML-based corpus formats such as TASX and NXT allow interoperability and the continued use of generic tools. It is now possible to annotate speech using speech analysis software such as Praat or ESPS/waves+, automatically calculate the structural relations between the tiers and finally query and process the data with the NITE tools. Existing phonetic corpora can thus be seamlessly integrated into structure oriented tree banks. This will increase the quality of linguistic resources by supporting queries of multi-level corpora which may provide evidence for the co-occurrence of phenomena on different levels of the annotation and confirm certain assumptions about the linguistic phenomena represented in the corpora.

One of the tasks to be carried out in the near future is the integration of the LeaP meta data into the TASX2NITE conversion. For some queries it would be interesting or necessary to be able to compare particular subsets of speakers such as speaker groups with different native languages or levels of competence.

6. References

[1] Bird, S.; Harrington, J., 2001. Speech annotation and corpus tools. *Speech Communication* 33, 1-4.

[2] Carletta, J.; McKelvie, D.; Isard A. (2002). Supporting linguistic annotation using XML and stylesheets. In *Readings in Corpus Linguistic*, G. Sampson; D. McCarthy, Continuum International.

[3] Cassidy, S.; Harrington, J., 2001. Multi-level annotation in the Emus speech database management system. *Speech Communication*, 61-77.

[4] Cassidy, S. 2002. XQuery as an Annotation Query Language: a Use Case Analysis, *Proceedings of LREC 2002*, Las Palmas, Spain.

[5] Erk, K.; Kowalski, A.; Padó, S.; Pinkal, M., 2003. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. Accepted at *ACL 2003*, Sapporo.

[6] Giegerich, H. 1992. *English Phonology*. Cambridge: Cambridge University Press.

[7] Grabe, E. 2002. Variation Adds to Prosodic Typology. In B. Bel & I. Marlien (eds.), *Proceedings of the Speech Prosody conference*. Aix-en-Provence: Laboratoire Parole et Langage, 127-132.

[8] Granger, S.; Hung, J.; Petch-Tyson, S. 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

[9] Grice, M.; Baumann, S.; Benzmüller, R. 2002. German Intonation in Autosegmental-Metrical Phonology. In *Prosodic Typology*, Jun, Sun-Ah (ed.), Oxford University Press.

[10] Gut, U.; Milde, J.-T. 2002. The Prosody of Nigerian English. In B. Bel & I. Marlien (eds.), *Proceedings of the Speech Prosody conference*. Aix-en-Provence: Laboratoire Parole et Langage, 367-370.

[11] Katz, H. 2003. *XQuery from the Experts: A Guide to the W3C XML Query Language*, Addison Wesley.

[12] Kay, M. 2001. *XSLT: Programmer's Reference, 2nd edition*, Wrox Press.

[13] Kipp, M., 2001. Anvil - A Generic Annotation Tool for Multimodal Dialogue, *Proceedings of Eurospeech, Aalborg*, 1367-1370.

[14] Li, A. 2002. Chinese Prosody and Prosodic Labeling of Spontaneous Speech. In B. Bel & I. Marlien (eds.), *Proceedings of the Speech Prosody conference*, Aix-en-Provence: Laboratoire Parole et Langage, 39-46.

[15] Marcus M.; Santorini B.; Marcinkiewicz M. A., 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19/2, 313-330

[16] Milde, J.-T.; Gut, U., 2001. The TASX-engine: an XML-based corpus database for time aligned language data, *IRCS Workshop on Linguistic Databases*, University of Pennsylvania, Philadelphia, USA.

[17] Milde, J.-T.; Gut. U., 2002. A Prosodic Corpus of Non-Native Speech. In B. Bel & I. Marlien (eds.), *Proceedings of the Speech Prosody conference*. Aix-en-Provence: Laboratoire Parole et Langage, 503-506.

[18] Mixdorff, H. 2002. Speech Technology, ToBI and Making Sense of Prosody. In B. Bel & I. Marlien (eds.), *Proceedings of the Speech Prosody conference*. Aix-en-Provence: Laboratoire Parole et Langage, 31-38.

[19] Peterson, G.; Lehiste, I. 1960. Duration of Syllable Nuclei in English. *Journal of the Acoustical Society of America* 32 (6), 693-703.

[20] Schiller A.; Teufel S.; Stöckert C.; Thielen C., 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. <http://www.ims.uni-stuttgart.de/projekte/corplex/german-tagsets.shtml>

[21] Schmid, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

[22] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J.; Hirschberg, J., 1992. ToBI: A standard for labeling English prosody. *Proceedings of the 1992 International Conference on Spoken Language Processing*, 867-870.

[23] Stirling, L., Fletcher, J., Mushin, I.; Wales. R., 2001. Representational issues in annotation: Using the Australian map task corpus to relate prosody and discourse structure. *Speech Communication* 33, 113-134.

[24] Wells, J.; Barry, W.; Grice, M.; Fourcin, A.; Gibbon, D., 1992. Standard computer-compatible transcription. *Stage Report Sen.3 SAM UCL-037*. University College London.

[25] Wightman, C. 2002. ToBI or not ToBI? In B. Bel & I. Marlien (eds.), *Proceedings of the Speech Prosody conference*. Aix-en-Provence: Laboratoire Parole et Langage, 25-30.