QUERY LANGUAGE FOR RESEARCH IN PHONETICS

Ulrich Heid, Andreas Mengel
Institut für maschinelle Sprachverarbeitung, University of Stuttgart,Germany

ABSTRACT

With the growing availability of spoken language corpora more and more data driven research in phonetics is possible. The downside of having huge speech corpora is that they have to be segmented and labeled, before they can be exploited. As labeling and annotation are time-consuming and costly, there is an interest in standardization which would support the exchange and reuse of labeled data. The MATE project proposes standards for an integrated and consistent multi-level annotation of speech and especially dialogue corpora. These proposals are based on the existing TEI standard (Text Encoding Initiative). All label information is represented in XML, thus there is a uniform representation of the different linguistic levels of description. This makes the implementation of tools easier and provides uniform access to the data, e.g. phonetic segmentation, prosodic labeling, grammatical annotation, dialogue acts classification, etc.

For the retrieval of information across multiple levels, a special query language and a query processor were developed. The query language was designed for the purpose of specifying linguistic items, contexts and constellations of phenomena to be found in spoken (dialogue) data. Basic concepts of this query language (called Q4M) are operators that let the user address both hierarchical (i.e. theory dependent) structures and physical (i.e. phenomenological) relations of linguistic objects.

The query processor is integrated into a software environment that allows the user to view results and to reformulate the query for further refinement and exploration of results. Thus, with the help of Q4M it will be easier, for example, to identify speech segments of variable length for the extraction and use in concatenative speech synthesis systems, or to investigate the interplay of speech acts and intonation and to test relevant hypotheses.

1. INTRODUCTION

In the field of phonetics, more and more speech corpora are produced. This is a great opportunity for both application oriented and basic research. Yet, the mere availability of recorded and phonetically segmented speech - be it manually or automatically produced - is only a first step towards the application of speech material.

As the object of description of phonetics is not only the physical structure of sequences of phones but also the functional aspects of speech, it is necessary to take other levels of description - e.g. syntax, semantics, pragmatics - into account to understand sources of and interactions in the rich structure of the acoustic manifestations of speech.

As a matter of fact the linguistic description of speech data is a costly enterprise, and only few institutions have the resources to label speech data on all relevant linguistic levels of description. Thus, an exchange of labeled data is desirable. Unfortunately, there are a number of different label formats used (e.g. [1,2]) which leads to a need for conversion software. Obviously, some standardization is required in this field. Moreover, so far, only few attempts have been made to propose a unified representation scheme for spoken language data covering several levels of linguistic description. Such an integrative approach at the same time serves the needs of exchange and of homogenous access to information from the different levels of description.

To develop a standard for the representation of linguistic information in spoken language data, a number of criteria have to be fulfilled. The respective formalism has to be applicable irrespectively of the level described and the theory used; open, to accommodate new descriptions; capable of linking information from different levels.

The MATE [1] project works towards a format for the uniform encoding of linguistic annotations; the representation formalism will be supported by guidelines for its effective use at five different levels of linguistic description (used as examples), and by portable software for the reading and writing, annotation and inspection, query and retrieval of data in this format.

The remainder of this paper will describe the format and the query language that supports the effective retrieval of annotated linguistic entities.

2. ENCODING

2.1 Existing standards

MATE has chosen the levels of prosody, morphosyntax, coreference, dialogue acts, and communication problems as examples of levels of linguistic annotation; of course other types of information may need to be encoded in spoken language data, too.

For each level of description, there is a tradition in speech and language processing, as regards theories and (application oriented) descriptive approaches, but also with respect to the formats (i.e. the syntax of the annotation formalism) used to express what the linguist has to say about a given stretch of speech. Well known application-dependent annotation formats (which sometimes combine theory and representation formalism) are xwaves/xlabel [1] or the PARTITUR format [2]. Similarly, there are theoretically inspired descriptive approaches for the levels of morphosyntax (e.g. the EAGLES standards proposal [3]), coreference, etc.

A formalism which is content-neutral and intended to encode any type of information, is XML.[4]. The Text Encoding Initiative, TEI [5], has come up with a series of (general) proposals for the encoding of layout- and content-related aspects of written language.

2.2 Requirements

Requirements for the coding of information in spoken language

Description	Example	Operators	Explanation
Comparison of elements by the			
values of their attributes			
to a string	($a.pos ~ "N")	~ !~	equals, does not equal
to a numerical value	($a.start < 0.2)	< <= > >= == !=	less, more, equal, not equal
relative to a other values			
as a string	($a.pos ~ $b.pos)	~ !~	equals, does not equal
as a numerical value	($a.end > $b.end)	< <= > >= == !=	less, more, equal, not equal
and a change	($a.f0 > $b.f0 * 2)	+ - * /	(mathematical operations)
position relative to other elements			
in a hierarchy	($a ^ $b)	^	is parent of
in a sequence	($a << $b)	, <<	is direct/any left neighbor of
related to time	($a [[$b)	% [[ ]] []][// @	(time relations cf. figure X)
membership of a set of			
elements	($a { $b)	{ !{ {}	is member, no member, join set
attribute values	($a.pos { $b.pos)	{ !{ {}	is member, no member, join set

Negation ("!") of single expressions and the combination of query expressions by logical operators ("&&" and "||") is also supported.

Figure 1. Overview of operators available in Q4M.

data are hard to stipulate, but there are at least some general minimal requirements. First of all, if a format is proposed as a standard, it should have been in use already for some time, thus there will be experience, expertise, and software available. Second, the grammar to be used should be universally applicable, i.e. as many of the phenomena and relations defined by the respective theories as possible should be representable by the standard: thus any information encoded this way will be parsable and interpretable.

2.3 XML

XML fulfills the requirements stated above. It uses entities (to encode, for example, linguistic objects) and properties attached to them. Entities are identified by an element name in brackets, e.g. <word>, properties can be defined <word start="0.23"/>. Elements can be nested, e.g. To represent a hierarchical structure:

```
<sentence><word>Hello!</word></sentence>
```

Elements can also be linked to one another, e.g. if they are held in separate files:

```
<word href="otherfile.xml#id(word_02)"/>
```

Since these general conventions are binding for all XML data, any XML parser and related applications can read and represent the elements, their attribute values and the relations among different elements encoded in a given document.

2.4 The MATE Coding Approach

The approach of the MATE project is to define sample annotation schemes for the individual linguistic levels chosen as example cases (prosody, morphosyntax, coreference, dialogue acts, communication problems), in a way that makes it possible to relate different entities from within one level of description, as well as entities belonging to different levels. For this purpose a number of XML guidelines are developed that instruct coders and developers on the encoding of speech and text data on one or more levels, so that they can be consistently processed. In part, existing schemes (such as, for example, ToBI for the functional aspects of prosody) have been re-expressed in XML, for the purpose of MATE encoding.

3. The MATE QUERY LANGUAGE Q4M

3.1 Purpose

The linguistic annotation of speech data is not a purpose in itself, but is an investment for later inspection and analysis of the data. In

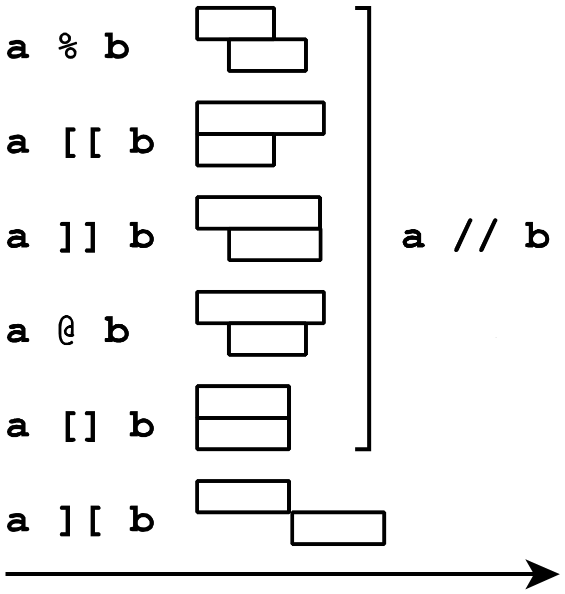


Figure 2. Time relations. The pairs of boxes represent linguistic objects and their relative extension in time.

order to be able to effectively carry out corpus-based research on spoken language data, the first step consists in providing all information in the same format. The query language is based on this assumption of a homogeneous encoding in XML, of annotations from all levels of description. Also, because of the XML format, different types of entities (elements) can be searched for (e.g. turns, sentences, chunks, words, etc.). The query language Q4M supports not only the search for entities from different levels of description, but also allows to include any structural relationships encoded in the XML documents in the set of constraints that define the query result.

Operators and operations available in Q4M are shown in figure 1. Particular attention has been paid to the relationships between segments with respect to the timeline: Figure 2 shows the kinds of constellations which are covered by time query constructs.

3.2 The Q4M Query Processor

The query is parsed, evaluated and transformed into a number of actions by accessing a database like representation of the structure of the XML files, e.g. those illustrated in figure 4.

```
word.xml:
<w  id = "w_01"  pos = "NE" >Flamsteed</w>
<w  id = "w_02"  pos = "VMFIN" >mußte</w>
<w  id = "w_03"  pos = "PPOSAT" >sein</w>
<w  id = "w_04"  pos = "NN" >Teleskop</w>
<w  id = "w_05"  pos = "ADV" >selbst</w>
<w  id = "w_06"  pos = "VVFIN" >kaufen</w>

pros.xml:
<tob id = "t_01"  type = "H*L"
  href = "word.xml#id(w_05)"/>
```

Figure 4. Sample corpus in XML.

3.3 Output Handling

Depending on the intended use, the output of a query may have to fulfill different needs: for example, it may just have to be displayed, or it may need to be 'piped' onwards to further processing (e.g. more refined queries, statistical analysis, etc.), or it may be fed back into the corpus by way of 'ad hoc labeling'.

First of all the data found matching the query specification should be returned. In the environment described here, there is no distinction between the context of a targeted element and the element itself. Instead, the whole set of query conditions is seen as the specification of a constellation of linguistic objects to be found. This makes even more sense since the information units that are needed to express the conditions are of various nature (e.g. may stretch over different time windows) and are represented on different linguistic levels. Another advantage of this approach is that elements of any level at any position can be specified within one of these constellations.

As queries of this kind might relate to quite complex situations, the user must be in a position to control the appropriateness of his query and its output. Therefore, the output of queries is a list of tuples of elements for which the conditions defined by the query are true. Also, results of sub-expressions can be inspected.

The output is represented in XML using href attributes that link the output to the original data (figure 5). The representation of the data by means of XML also includes reference to the query expression itself. As the output of a query is represented in XML, it can serve as a new document to be searched or as a source of refinement and documentation of queries executed.

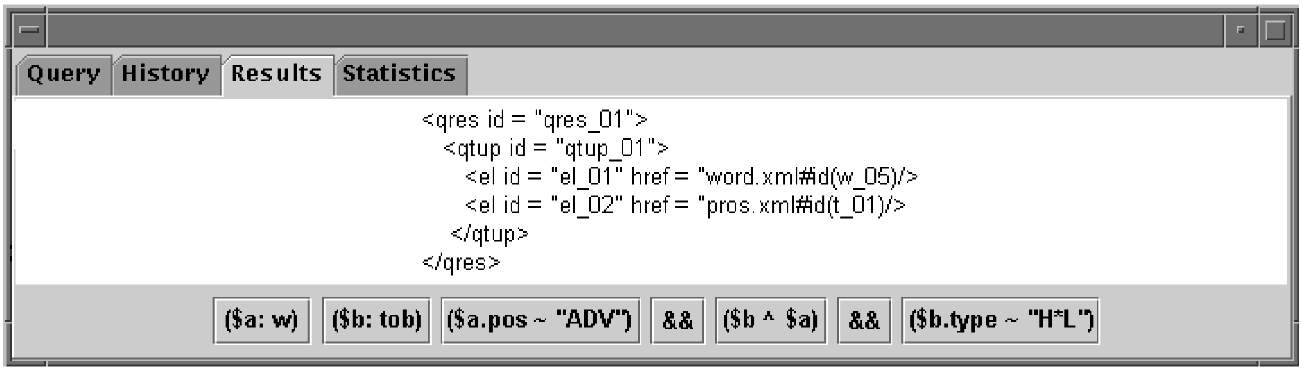
3.3 Application of Results

When producing linguistically marked-up databases, the query language can be used for inspection of the data: selective retrieval makes it easier to identify inconsistencies in the annotations. As the output of queries constitutes one or more new documents of XML, this can be used to produce label data with phenomena that are defined by the constellation of previously tagged entities.

Within basic research, the use of Q4M and its environment can improve the validation of hypotheses: They can be defined as queries, tested against the output and confirmed or reformulated. Also, the integration of phonetics and other levels of linguistic theory can be put forward: In speech synthesis, Q4M and the MATE software environment can be used to identify appropriate portions of the speech signal when optimizing the unit selection approach [6].

3.4 Example

The following example shows how prosodic data can contribute to the disambiguation of annotations from other levels, in this case



```
<qres id = "qres_01">
  <qtup id = "qtup_01">
    <el id = "el_01" href = "word.xml#id(w_05)/>
    <el id = "el_02" href = "pros.xml#id(t_01)/>
  </qtup>
</qres>
```

Figure 5. Query expression result in XML representation.

word class annotations. The German word selbst has several functions, analysed usually as readings with different word class properties. In a sentence like (1), it is a reflexive pronoun (engl. oneself), whereas it is a focus particle (roughly equivalent to even) in (2). This homography typically comes with intonation differences visible in the intermediate context:

(1) *Damit schadet man sich selbst und seinem Nächsten.*
(2) *Flamsted mußte sein Teleskop selbst kaufen.*

The category difference of selbst is often very hard for stochastic part-of-speech taggers to identify, and tools for e.g. the construction of language models may in many cases not get the right readings. An example is the incorrect POS-assignment in (3):

(3)	*Flamsteed*	*mußte*	*sein*	*Teleskop*	*selbst*	*kaufen.*
	NE	VMFIN	PPOSAT	NN	ADV!	VVINF
	Flamsteed	had to	his	telescope	himself	buy

In a corpus which contains prosodic annotations (in our example, ToBI labels), we can retrieve all constellation of possibly mistagged instances of *selbst* and one of the two intonation patterns. So, if the data are represented as shown in figure 3 all cases with the typical reflexive pronoun intonation (such as in (3)) can be retrieved with the query expression in figure 6.

```
($a: w) ($b: tob);
($a.pos ~ "ADV") && ($b ^ $a) && ($b.type ~ "H*L")
```

Define variable $a *to refer to* w *elements and define* $b *to refer to* tob *elements.*
Find those w *elements which have the orthographical representation of* selbst, *a* POS *value of* ADV, *and which are refered to by a* TobI *label that is classified as* H*L.

Figure 6. Example query.

The same is possible, of course, for all cases of selbst as a focus particle.

4. CONCLUSION

In this paper an environment for the consistent and uniform annotation, representation, and retrieval of entities of linguistic databases has been described. The query language Q4M as part of this scenario can serve to access and gain insight into multilayered information in XML tagged corpora. It is the hope that this contribution will serve to integrate phonetic processes with other levels of linguistic description, to retrieve data on their interplay and to thereby gain new insights into the interdepencies between phonetic and non-phonetic aspects of linguistics.

References

[1] World Wide Web: http://www.entropic.com/products/esps/esps.html
[2] Schiel, F., Burger, S., Geumann, A., and Weilhammer, K. 1998. The Partitur Format at BAS. *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, Part II, 1295-1301.
[3] World Wide Web: http://www.ilc.pi.cnr.it/EAGLES/home.html
[4] World Wide Web: http://www.w3.org/XML
[5] Sperberg-McQueen, C.M. and Burnard, L. (Eds.). 1994. *Guidelines for Electronic Text Encoding and Interchange.* TEI P3. Text Encoding Initiative. ACH, ACL, ALLC, Chicago, Oxford.
[6] Mengel, A. and Heid, U. 1999. Query Language for Access to Speech Corpora. *Forum Acousticum 1999,* Berlin.

[1] MATE is a project in the Language Engineering Programme of DG XIII E, Luxemburg (LE4-8388); part of the work described here was funded in the framework of the MATE project.