ENHANCING REUSABILITY OF SPEECH CORPORA BY HYPERLINKED QUERY OUTPUT

Andreas Mengel, Ulrich Heid
IMS Stuttgart University
Azenbergstraße 12, 70190 Stuttgart, Germany
{mengel,heid}@ims.uni-stuttgart.de
http://www.ims.uni-stuttgart.de

ABSTRACT

In speech technology more and more databases of spoken language are becoming available. For research the availability of these data offers the possibility to study huge corpora. Apart from the fact that these corpora may be represented in different formats, it is sometimes difficult to relate annotations of one corpus to those of another corpus. This contribution argues for a representation of information in speech corpora that allows for the integrated representation of information on various levels of description in XML. Secondly, the study of huge amounts of speech data requires adequate retrieval mechanisms. A query architecture is described that allows for the retrieval of encoded entities by specifying their properties or various relations to other entities. The output of the query processor is represented in XML and thus can be used for further queries or a new level of description. The work presented here is part of the results of the MATE project (http://mate.mip.ou.dk).

1. INTRODUCTION

The number of corpora of spoken language has considerably increased over the last few years. Yet, the sheer mass of data does not guarantee accessibility and usability of the material. Two more steps are necessary to achieve this: **description** of speech and language of the data and **access software** for retrieval and exploration of relevant annotated subsets. The linguistic description and annotation of speech data is a costly task. In order for this investment to pay off, access to annotated data must be uniform and as easy as possible: consequently, both an adequate representation formalism and retrieval tools operating on that formalism are needed.

Description of speech and language data: The MATE project (a project in the Linguistic Engineering programme of DG XIII E of the European Commission) is working towards proposals for uniform encoding procedures and for annotations based on an extension of the TEI guidelines using XML. The descriptive levels of prosody, morpho-syntax, coreference, dialogue acts and communication problems in dialogue systems are used as an exemplification of the approach: for these, annotations are being worked out an tested on several European languages; the markup framework is generic, such that linguistic descriptions from other levels can be added without difficulties. Reusability depends crucially on documentation, it involves a reinterpretation of annotations found in a corpus, to interpret these, information on the underlying classification criteria is crucial.

Access software: For any purpose, be it research or application oriented, flexible access and retrieval tools for massively annotated speech corpora are needed: anything that can be annotated in the data must be retrievable, and any combination of encoded information, text and speech must be searchable. However, everything that can be searched for and is found has also to be accompanied by equally structured output. In MATE, a query processor (called Q4M) for XML-annotated speech and language corpora has been developed which satisfies the requirements stated above. Query results are again XML documents, where the retrieved elements are linked to both the query expression and the documents they originate from. As a consequence, query results may themselves be further searched, and 'cascaded' queries be constructed. Moreover, units of different size (words, word sequences, phrases, sentences, turns) and annotations from all levels can be combined. This is particularly useful for analyses of data which can only be derived from the co-occurrence of phenomena typically described at different levels, for example the identification of dialogue acts or the analysis of relative clauses (restrictive vs. explicative), as related with prosody.

2. ANNOTATION

Standards for the description of speech and language data require a representation format that is independent from the phenomenological aspect described (signal, perception, or linguistic function), from level of description (e.g., prosody, syntax, or dialogue acts) or theoretical approach used to understand the phenomena. XML [1] is such a standard which allows for the definition of entities – called *elements* – their property dimensions – *attributes* and *values* – and requires only one kind of parser for all markup information represented in this format.

Different aspects of description of language data are not independent of each other and will mostly be produced in an iterative fashion, for example, the description of

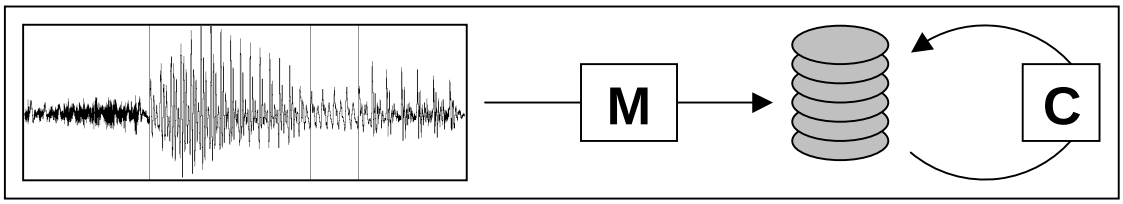

Figure 1: Two kinds of annotation processes. M: Annotation by measurement (human perception or physical measurement of speech signal), C: Annotation by categorisation (human assignment or algorithmic processing of existing annotation data).

intonation requires inspection of fundamental frequency information, and for morpho-syntactic annotation a segmentation of utterances into words is indispensable. Thus, the process of annotation can be described as implying two different kinds of processing, namely measurement and categorisation. *Measurement* is the direct processing of physical aspects of the speech signal itself, *categorisation* is any further processing of the data measured or categorised (Figure 1).

For an integrative representation of speech data annotation that spans different levels of description and represents their relatedness, special means are required. Obviously, the immediate neighbourhood of entities is reflected by the sequence of their annotation. The standard means, in XML, of representing hierarchically organised entities – e.g. sentences and words - is a nesting of elements:

```
<sent>
<word>Die</word>
<word>Sonne</word>
<word>lacht</word>
</sent>
```

Another mechanism to represent the hierarchical relations of elements in XML is the use of hyperlinks (href) (Figure 2) [2].

This second approach of linking the description of phenomena encodes the relatedness of both their theoretical status and the procedure of annotation. In most cases where many related levels of descriptions are involved, the use of href attributes is more practically as the (iterative) processing of information of different files is less error prone. Yet, the fact that different items are linked can also be used for an effective representation: As a subset of properties of elements are constitutive for the linking of information – in the case of word and sentence annotation it is the time information of the outer boundaries – these are properties which can also be inferred via the href links and need only be represented on one level.

Thus use of href links in the representation of annotation data is also a good means of non-redundant representation of information, requiring that time information, speaker information, etc. need only be specified on the first – i.e. lowest - level of annotation. Any information that is applicable for a given level of description and is not explicitly specified can be retrieved from lower levels of description. This will apply to time information, as well as to the language identifier, the speaker, and other aspects. Note that the reference to other levels of description need not always imply a part-whole relationship as in the case of sentences and words: The linking between any two levels of description may in most cases only reflect the order in which the annotations were added to the text; an example is fundamental frequency information and prosodic labelling.

The use of href attributes is not limited to the inference within annotated corpora. Parts of the information anno-

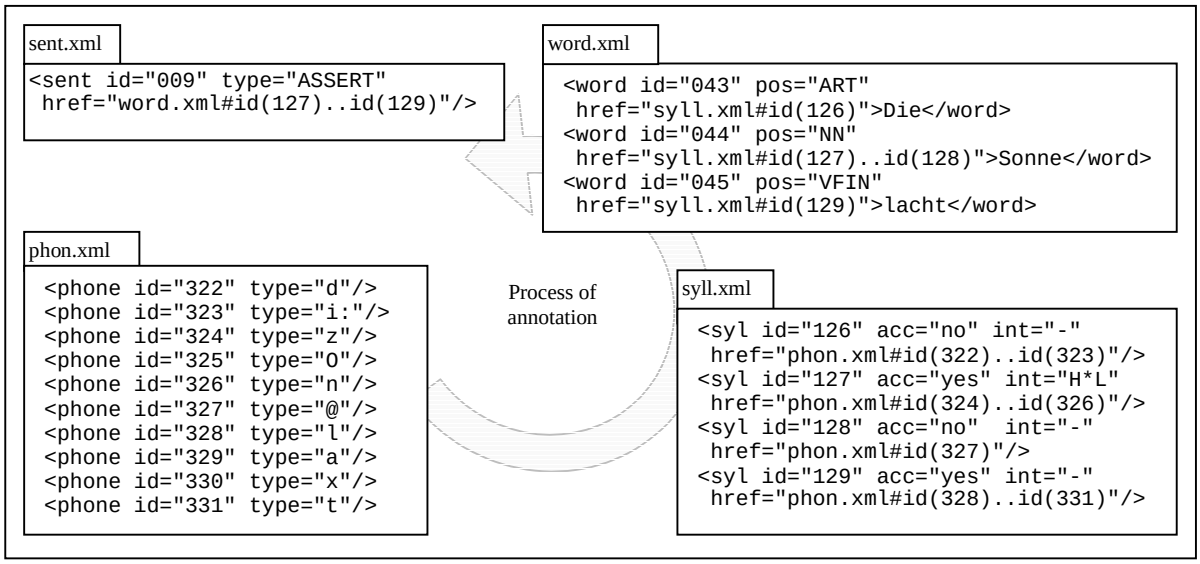

Figure 2. Iterative process of annotation and hyperlinking: phones, syllables, words, and sentences.

tated in language data is inferable from lexical resources. Also, if dialogues are annotated with communication problem markup, the problematic areas of the dialogues may be related with a database of cooperativity guidelines, to point to the guidelines not observed in the given problem case. Therefore, it seems plausible to provide linking mechanisms to other resources outside the corpus.

To summarise, annotation of speech data and its representation is mandatory for accessing, manipulating, and processing the speech signal itself or building up knowledge resources. Most annotation is based on existing annotation and information items may be identical across annotation items and may thus be used for inference processes. Some information should be encoded only once in a corpus. This will enhance the consistency and support the maintenance of the data, as changes that apply to many items only have to be made once, i.e. at the items that hold the respective information to be altered. Yet, even in an ideal situation, any annotation will only represent the knowledge about the tagged phenomena and their relations that can be described and are understood by the actual theoretical apparatus.

3. QUERY

Annotation data must be accessible. In MATE, a query language and a query processor have been developed, that allow for the retrieval of annotated data in many respects. Not only can elements with special properties be specified, but also their relation to other elements, hierarchical relations etc. Figure 3 gives an overview of the query concepts of the query language Q4M.

If the input documents are the XML files shown in figure 2, a possible query could be:

*Find all [O] phones in accentable syllables that contain a ToBI label of type H*L in those nouns that are located in sentences that are assertions.*

This query would be represented by the following query expression, which consists of a variable declaration part and a query specification part:

```
($p:phone)($sy:syl)($w:word)
($st:sent);

($p.type~"O")&&($sy^$p)&&
($sy.acc~"yes")&&($sy.int~"H*L")&&
($w^$sy)&&($w.pos~"NN")&&($st^$w)&&
($st.type~"ASSERT")
```

The output of queries are again XML documents which provide links to all those constellations of elements which satisfy the conditions stated in the query expression. The output also includes the original query string.

```
<qures id="qures_001"
 quexpr="($p:phone)($sy:syl)
         ($w:word)($st:sent);
         ($p.type~"O")&&($sy^$p)&&
         ($sy.acc~"yes")&&
         ($sy.int~"H*L")&&($w^$sy)&&
         ($w.pos~"NN")&&($st^$w)&&
         ($st.type~"ASSERT")">
  <qutup id="qutup_1">
    <el id="el_001"
     href="phon.xml#id(phone_325)"
     refvar="p"/>
    <el id="el_002"
     href="syll.xml#id(syl_127)”
     refvar="sy"/>
    <el id="el_003"
     href="word.xml#id(word_044)"
     refvar="w"/>
    <el id="el_004"
     href="sent.xml#id(sent_009)"
     refvar="st"/>
  </qutup>
</qures>
```

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 << $b)`	`, <<`	is direct/any left neighbour of
related to time	`($a [[ $b)`	`% [[ ]] [] ][ // @`	(time relations)
membership of a set of			
elements	`($a { $b)`	`{ !{ {}`	is member, no member, join sets
attribute values	`($a.pos { $b.pos)`	`{ !{ {}`	is member, no member, join sets

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

Figure 3. Operations supported by the query language Q4M.

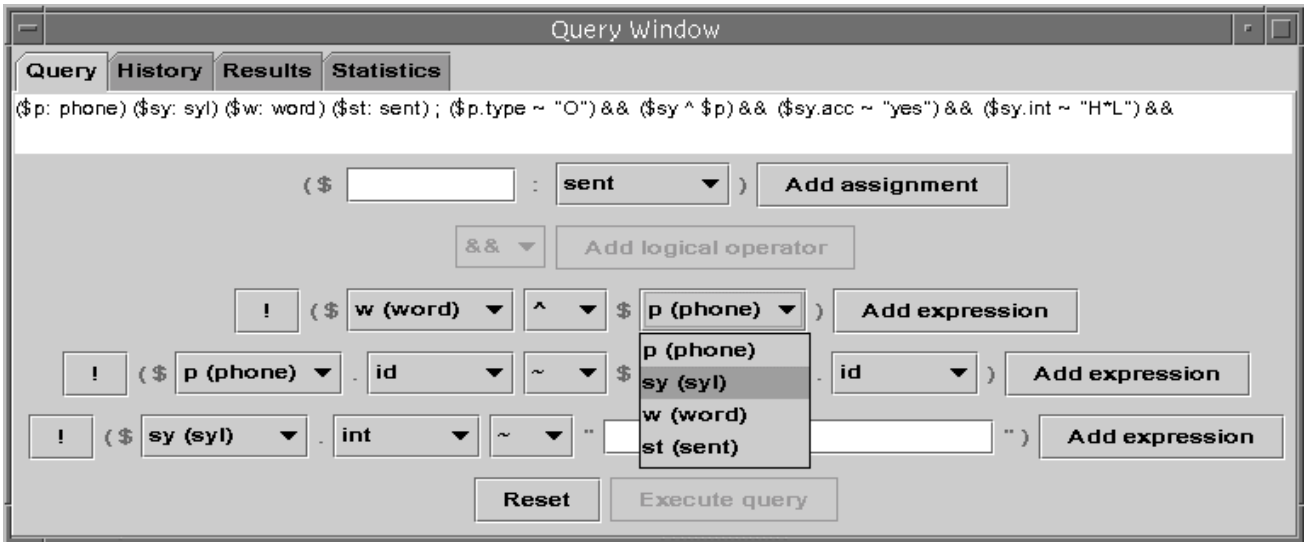

Figure 4. Interactive query formulation support interface.

The <el> elements have an href attribute that points to the elements found to match the query; by the refvar attribute, the sub-expressions of the query expression (defined in the qexpr attribute) can be identified that specify conditions for this entity.

By providing href links to the element tuples that satisfy a given condition stated in a query expression, access to the items found is guaranteed. Thus, first of all it is possible to refine the query expression or inspect the data found and their context. Secondly, the output of queries – being represented in XML – can be used as a new kind of annotation. If – as stated above – the annotation of speech data in XML represents the current state of theories available for linguistic phenomena and their interrelationships, then the output of queries can be used to formally describe any relation not explicitly encoded in the annotation data. These descriptions might be arbitrary as there are many meaningless relations among different elements. Yet, if queries are used for the exploration and refinement of hypotheses, the output will exactly serve five requirements needed for this task: The precise description of the relations among the entities that are constitutive for the particular phenomenon; the possibility to inspect the environment of the result items by hyperlinks; the iterative refinement of queries or formulation of queries that query queries (the interactive support tool (cf. figure 4) also allows to inspect intermediate results of the query); the use of query output as a new level of annotation; and the documentation of the queries in XML.

4. SOFTWARE ENVIRONMENT

The query processor is integrated in a software architecture (written in JAVA) that was developed for the support of human annotation and inspection of XML encoded dialogue data. The query processor interacts with the internal representation of the system [3] that reads the XML encoded data into memory and represents all relations specified. This representation of the annotation information is accessed by the query component during interactive query formulation process and data retrieval. As the output of the query is XML, that refers to the elements found, this new information is also read into the representation unit of the software thus adding to the existing annotation.

5. CONCLUSION

In this contribution we argued for the provision of links between information entities in annotated speech corpora. The encoding environment for this is XML, as XML is seen as the most powerful and most general encoding standard available at the moment. The means for relating annotations of any entities are href attributes that specify the ID of the entities referred to and the files they are located in. The linking of information within and across annotated documents is useful for a consistent encoding and maintenance of information that implies the distributed location of properties and inference of information by the exploitation of the link structure. The same means are used for the linking of query output of the query processor Q4M. By providing links to tuples of elements that satisfy the constraints of a given query expression, the output of queries serves as access document to these elements, as a formal description and as documentation of the query process.

6. REFERENCES

[1]World Wide Web: http://www.w3.org/XML

[2] Isard, A., McKelvie, D. and Thompson, H.S (1998), Towards a Minimal Standard for Dialogue Transcripts: A New Sgml Architecture for the HCRC Map Task Corpus. *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP98*, Sydney.

[3] Isard, A., McKelvie, D., Cappelli, B., Dybkjær, L., Evert, S., Fitschen, A., Heid, U., Kipp, M., Klein, M., Mengel, A., Møller, M.B. and Reithinger, N. (1998), Specification of workbench architecture. MATE Deliverable D3.1.