New features in Spoken Language Search Hawk (SpLaSH): Query Language and Query Sequence

Sara Romano[1], Francesco Cutugno[2]
[1]SECLAB, Department of Computer and System Science
[2]LUSI-Lab, Department of Physical Sciences
Federico II University, Naples, Italy
sararomano@na.infn.it, cutugno@na.infn.it

Abstract

In this work we present further development of the SpLaSH (Spoken Language Search Hawk) project. SpLaSH implements a data model for annotated speech corpora integrated with textual markup (i.e. POS tagging, syntax, pragmatics) including a toolkit used to perform complex queries across speech and text labels. The integration of time aligned annotations (TMA), represented making use of Annotation Graphs, with text aligned ones (TXA), stored in generic XML files, are provided by a data structure, the Connector Frame, acting as table-look-up linking temporal data to words in the text. SpLaSH imposes a very limited number of constraints to the data model design, allowing the integration of annotations developed separately within the same dataset and without any relative dependency. It also provides a GUI allowing three types of queries: simple query on TXA or TMA structures, sequence query on TMA structure and cross query on both TXA and TMA integrated structures. In this work new SpLaSH features will be presented: SpLaSH Query Language (SpLaSHQL) and Query Sequence.

1. Introduction

The production of language corpora is constantly evolving. In recent years, many linguistic corpora have been enriched with the addition of different annotation levels. As it is well known, spoken language corpora can contain both acoustic-temporal and textual-transcriptional levels. The acoustic-temporal levels are referred to annotations describing the acoustic properties of speech signal (intonation, phonemes, etc.). Being strictly dependent on the signal and thus on the time, these types of annotations are defined as time-aligned annotations (TMA). At the same time, the textual levels are referred to annotations resulting from the analysis of transcriptions (syntax, pragmatics, morphosyntax, etc.) and therefore they are defined as text-aligned annotations (TXA). Additionally, annotation levels can be related together by hierarchical relations expressed both among annotations (inter-level) and within a given annotation (intra-level). An example of intra-level hierarchy is the syntactic textual parsing where sentences are divided into phrases which furtherly are subdivided into smaller units. Moreover a corpus may be characterized by multiple hierarchies that may or may not share some levels. Linguistic knowledge representation makes in some cases difficult to define hierarchies that should be verified in all circumstances (Bird & Harrington, 2001). The passage from annotated datasets to integrated systems for information retrieval on these data requires the generation of specific databases and relative search tools. Furthermore, a lack of agreement on the storage format for the linguistic annotations leads to a further problem concerning the reusability of linguistic corpora. In fact it often happens that tools developed within a given project can not be reused. As a consequence, the integration of data coming from different sources requires additional efforts to transform a corpus storage format in another. For this reason, general purpose systems for the managing different annotation standards with multiple hierarchies have been developed. Generally these systems accept as input several formats of annotation and return a database that a user can search in by means of specific tools.

EMU Speech Database system (Cassidy & Harrington, 2001) and NITE XML toolkit (Carletta et al., 2005) are some of the most representative examples of these applications. Recently, we have presented SpLaSH (Spoken Language Search Hawk) a new general purpose system for multilevel linguistic corpora management (Romano et al., 2009). In SpLaSH data coming from different corpora are allowed and linguistic annotations belonging both to TMA and TXA categories are integrated. Differently from the EMU system, in SpLaSH no fixed hierarchies among the annotation levels are imposed; our system considers only those implicitly defined in the data model as it is based on the idea that each level could be obtained independently from the others. Differently from the NITE toolkit, in SpLaSH no metadata files are used to describe data structures hence no human intervention is needed to define internal organization of linguistic resources. In this work new SpLaSH features will be presented. As it will be shown later in this paper, we will present the following innovations: SpLaSH Query Language (SpLaSHQL) and Query Sequence.

2. SpLaSH Data Model

SpLaSH encodes TMA annotations through Annotation Graphs (AG) (Bird & Lieberman, 2001). Annotation Graphs are a descriptive model able to embody the main annotation formats (like TIMIT (Garofalo et al., 1993), Praat TextGrid (http://www.fon.hum.uva.nl/praat/), Partitur (Schiel et al., 1998)) and can be considered as a unifying standard in principle applicable to any speech corpus. As it is shown in the Figure 1, Annotation Graphs

are data structures including temporal references (represented by nodes in an oriented graph) anchored to the signal while the left to right oriented graph arcs are labeled with the couple of data annotation-type/value.

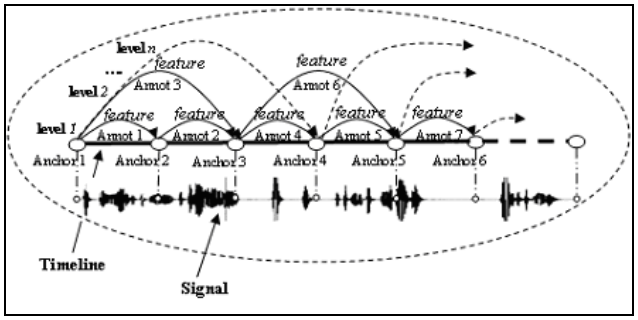

Figure 1: Annotation Graph data structure

The Annotation Graphs representing a speech corpus, are organized in a data structure called AGSet. SpLaSH implements TMA annotations coded as Annotation Graphs according to the standard XML native database format originally proposed by the AG's authors. Consequently, all the formal definitions proposed by the authors are respected.

TXA data can be recursive and, as already observed above, their descriptive elements included in the annotation system can be structured in a hierarchical fashion.

For this reason XML is considered the ideal instrument for these type of annotations too, as, in this way, it is possible to organize annotation elements in a tree structure, in which, if necessary, the sequential nature of the text, related to temporal development of speech units, is included in the organization of the leaves (see Figure 2 and Figure 3).

The usage of XML as a storage format ensures great freedom in the definition of the specific annotation system. In order to preserve this freedom, in SpLaSH the number and the nature of constraints imposed to the formal definition of the TXA data is very limited and the most relevant one requires that the transcribed text must be linearly represented at the level of the tree leaf. With reference to fig. 2 it means that strings *text1…text6* appear in the same order as they appear in the original text (or, alternatively, it must be possible, by means of an indexing procedure, to reconstruct the original sequence).

The integration of TMA and TXA data represents the main aim in SpLaSH. Both datasets are physically implemented in XML files. Furthermore TXA annotations values are stored in XML tree leaves while TMA ones are stored as arc labels (that are represented as XML leaves). In order to allow the integration of TMA and TXA annotation classes in an unique structure, we introduce one simple constraint on these annotation classes: *"TXA annotation values in the leaf must coincide with at least one level of the TMA annotations"*.

2.1 Connector Frame

The constraint defined above, leads to the definition of a new structure named Connector Frame (CF) that acts as interface between TXA and TMA annotation classes (see Figure 4).

The Connector Frame is also coded by an XML file and contains references to nodes belonging both to the TMA and TXA structures to create a whole structure. Essentially the CF has the form of a tree with a root, a child level containing ID-values of TXA nodes that are fathers of textual leaves and finally a level containing ID-values of corresponding TMA values. The integration process allows TXA nodes to inherit the temporal relationships from the TMA levels and allows a user to perform sophisticated analysis on such linguistic data.

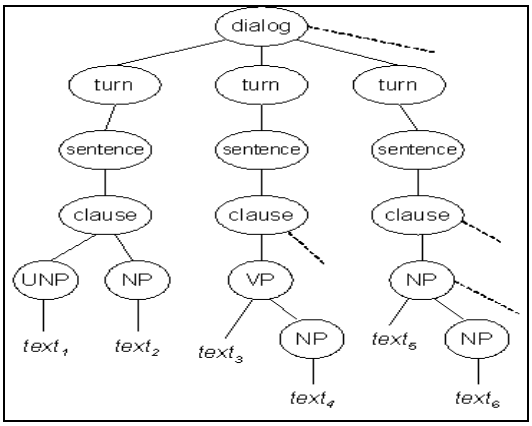

Figure 2: An example of the TXA logical representation: the case of syntax treebanks.

```
<?xml version="1.0" encoding="UTF-8"?>
<dialog dialog_id="DGmtB03p">
  <turn turnid="p1G#1" compl="f">
    <sentence uni="t" n_of_clauses="1">
      <clause type="m">
        <UNP>text1</UNP>
        <NP>text2</NP>
        ... .
      </clause>
    </sentence>
  </turn>
  <turn turnid="p1G#2" compl="f">
    <sentence uni="t" n_of_clauses="1">
      <clause type="m">
        <VP>
          text3
          <NP>text4</NP>
        </VP>
        ... ..
      </clause>
    </sentence>
  </turn>
  <turn turnid="p1G#3" compl="f">
    <sentence uni="t" n_of_clauses="1">
      <clause type="m">
        <NP>
          text5
          <NP>text6</NP>
          ....
        </NP>
        ... ..
      </clause>
    </sentence>
  </turn>
</dialog>
```

Figure 3: XML as storage format for TXA of Figure 2

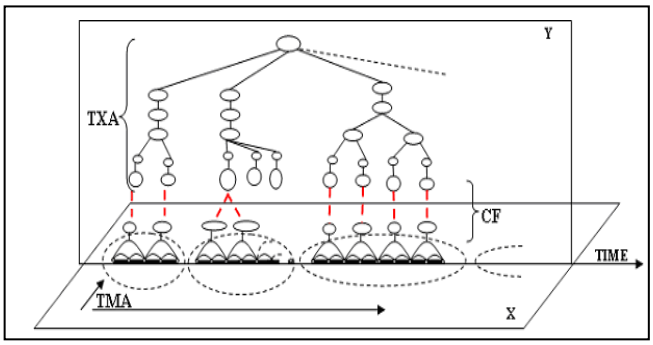

Figure 4: TXA and TMA integrated structure

In the Figure 5 a sketch of connector frame structure is given.

```
<listOfNodes>
    <node idnode= "idNodo">
        <arch idArch="id1"> text </arch>
        <arch idArch="id2"> text </arch>
        <arch idArch="id3"> text </arch>
    </node>
</listOfNodes>
```

Figure 5: Connector Frame

3. Query Language

In order to enable information retrieval on linguistic data belonging to the SpLaSH data model, we have formalized the SpLaSHQL query language. SpLaSHQL is based on a set of specific algebraic operators aimed at the semantic definition of the queries performed on the TMA and TXA integrated datasets. Currently a subset of queries – i.e. those that are more interesting for the linguistic research community - have been implemented using the XQuery (http://www.w3.org/TR/xquery) language (and of course XPath (http://www.w3.org/TR/xpath)). XQuery is a language that allows correct modeling of the sequential and hierarchical features in linguistic data (Cassidy, 2002). Suppose we want to perform the following query: *"Select all verbal phrases in the syntactical (TXA) annotation which are coincident with a (TMA) tone unit"*. Using SpLaSHQL this query is expressed by the following expression (Figure 6):

```
SELECT %vp
FROM   %vp ← doc(syntax.xml)//VP,
       #tu ← doc(ag.xml)//Annotation
WHERE  #tu/@type= "tu" AND (%vp COINCIDENT #tu)
```

Figure 6: An example of query expressed in SpLaSHQL

SpLaSHQL commands partly recall the SQL format with some XPath insertions used to filter the nodes in the XML documents.

Node filtering assumes different properties in the two types annotation: SpLaSHQL uses the '%' prefix to express variables connected to TXA nodes, while '#' is the prefix for variables indicating TMA nodes. The 'WHERE' clause in Figure 6 processes temporal constraints. In this case, the 'COINCIDENT' operator, that, in principle, can be used only on TMA nodes, extends its domain on TXA nodes that inherited temporal lables by means of the Connector Frame.
This query returns a node list made of couples of tma and txa objects satisfying the request expressed in figure 6. Thanks to the facilities offered by XQuery the output is redirected to an XML file having the tag <result> as root (see Figure 7).

```
<result>
    <VP>
    <arch idArch="id1" type="TU"> text </arch>
    <VP>
    <arch idArch="id2" type="TU"> text </arch>
</result>
```

Figure 7: Formal structure of output deriving by the query expressed in Fig. 6

To express such a request using XQuery code, would produce a huge amount of code lines and reduce the access to the query system less easy for not computer programming skilled linguists. SpLaSHQL constitutes a middleware between a high-level query interface available in SpLaSH and the relative XQuery implementation below.

4. Query Sequence

Let us recall that, sequential patterns represent a common requirement in linguistic corpora features research. Sequences queries on TMA structures are used to express the transitive closure over the arcs of annotation graphs. Query Sequence has been implemented in XQuery. The sequence retrieval function is based on an algorithm that accepts as inputs an AGSet *AGS*, an annotation level *L* and a target sequence *T*. The target sequence specifies a subsequence of contiguous annotations and is composed by a set of strings and the symbol '*' used to specify a contiguous length independent sequence of annotations. For example, the target sequence *T = word1 * word2* represents all the sequences of strings that begin with *word1* end with *word2* and contains any contiguous sequence of strings between them. No length limitations are imposed on the target sequence so it is possible to perform much more complex queries. An example of query sequence expressed with SpLaSHQL is shown in Figure 8. Such query returns all sequences of TMA word annotations that start with the word 'la' and ends with the word 'casa'. In Figure 9 the output deriving by the query expressed in Figure 8 is shown.

```
SELECT #w
FROM   #w ← doc(ag.xml)//Annotation[@type="wrd"]
WHERE  [#w//text()="la"].[#w]*. [#w//text()="casa"]
```

Figure 8: An example of a sequence query expressed in SpLaSHQL

```
<result>
    <sequence>
        <arch idArch="id1" type="wrd"></arch>
        <arch idArch="idn" type="wrd "></arch>
    </sequence>
    <sequence>
        <arch idArch="id1" type=" wrd"></arch>
        <arch idArch="id2" type=" wrd"></arch>
        <arch idArch="id3" type=" wrd"></arch>
        <arch idArch="idk" type=" wrd"></arch>
    </sequence>
</result>
```

Figure 9: Formal structure of output deriving by the query expressed in Fig. 8

5. Conclusions

To support end users work, SpLaSH include a GUI for query generation on data belonging to the SpLaSH data model (Romano et al., 2009). Three types of queries are allowed: simple query on TMA or TXA structures, sequence query on TMA structure and cross query on both TXA and TMA integrated structures. The GUI's underlying engine is implemented by XPath and XQuery code. Simple queries on TXA structure are based on XPath language while simple queries on TMA structure, sequence queries and cross queries are based on XQuery operators. Each class of query has its own graphical interface containing several components to facilitate the query generation.

Thus SpLaSH presents interesting innovations in the linguistic general purpose systems developing area. Our system imposes a very limited number of constraints to the data model design, allowing the integration of annotations developed separately within the same dataset and without any relative dependency. The graphical interfaces are designed to guide users to compose queries on the data model. Being a metalanguage that emphasizes simplicity, generality, and usability over the web, the choice of XML as the storage format for linguistic annotations, leads to improving the data reusability. The next step is to implement the SpLaSHQL query language in order to allow expert users to define new queries according to their needs.

Splash is an open source project available at http://s2snaples.fisica.unina.it/splash, under GNU-Public license.

6. References

Bird, S. and Lieberman, M. (2001). A formal framework for linguistic annotation. *Speech Commun.* 33, Issues 1-2, Pages 23-60.

Bird, S. Harrington, H. (2001). Speech annotation and corpus tools. *Speech Commun.* 33, Issues 1-2, Pages 1-4, Elsevier.

Carletta, J., Evert, S., Heid, U. and Kilgour, J. (2005). The NITE XML Toolkit: data model and query language. *Language Resources and Evaluation Journal*, Pages 313-334.

Cassidy, S. Harrington, J. (2001). Multi-level annotation in the Emu speech database management system. *Speech Commun.* 33, Issue 1-2, Pages 61-77, Elsevier.

Cassidy, S. (2002). XQuery as an Annotation Query Language: a Use Case Analysis, *Proceedings of 3rd Language Resources and Evaluation Conference* (LREC).

Garofalo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. and Dahlgren, N.L. (1993). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. *Technical Report NISTIR 4930*, NIST.

Romano, S., Cecere, E., Cutugno, F. (2009). SpLaSH (Spoken Language Search Hawk): integrating time-aligned with text-aligned annotations. *Proceedings of Interspeech*.

Schiel, F., Burger, S., Geumann, A., and Weilhammer, K. (1998). The Partitur Format at BAS. *In Proceedings of the First International Conference on Language Resources and Evaluation*, Pages 1295-1301.