SpLaSH (Spoken Language Search Hawk): integrating time-aligned with text-aligned annotations

Sara Romano, Elvio Cecere, Francesco Cutugno

Natural Language Processing group
Department of Physical Sciences, Federico II University, Naples, Italy
sararomano@na.infn.it, elvio.cecere@gmail.com, cutugno@na.infn.it

Abstract

In this work we present SpLaSH (Spoken Language Search Hawk), a toolkit used to perform complex queries on spoken language corpora. In SpLaSH, tools for the integration of time aligned annotations (TMA), by means of annotation graphs, with text aligned ones (TXA), by means of generic XML files, are provided. 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. Index Terms: search tools, speech corpora, time-aligned annotations, text-aligned annotations.

1. Introduction

In recent years a certain number of internationally available speech corpora have been enriched with many levels of annotation both at acoustic-temporal and textual (transcriptional) level. Annotation levels may be organized hierarchically. A corpus may be characterized by multiple hierarchies that may or may not have some levels in common. Tools developed within a given project rarely have been reused and the integration of data coming from different sources has required significant additional work, largely due to lack of agreement on the ideal storage format for the linguistic annotations. At the same time, general purpose systems for the management of different annotation standards with multiple hierarchies have been developed. These accept annotated files as input and return special format databases that a user can search by means of specific tools [1]. EMU Speech Database System [2] is a representative example of these applications: it is one of the first general purpose systems created for the management of heavily cross-annotated data. In EMU, the data are organized in tokens that represent some convenient unit of analysis such as words, phonemes or sentences. There are two types of tokens: events and segments. Querying the system entails retrieving events, segments or sequences of segments by navigating into a hierarchical tree that rigidly structures relations among levels of available annotation.

In EMU, temporal labels are associated with only one annotation level; all other levels inherit time information by means of structure of the hierarchy. This solution does not facilitate the reuse of limited subsections of the corpus and introduces a principle of temporal dependency between annotation levels which is not always acceptable. Another recent general purpose system is NITE XML toolkit [3]. NITE XML includes two kinds of relations between annotation levels: one organizes data hierarchically, the other defines ontologies for the description of more complex data. Thus, in contrast with EMU, NITE allows the presence of temporal references on more than one level. It permits multiple hierarchies but imposes the use of an ontology, stored into a metadata file, describing and fixing their structure. Consequently each change in the data structure requires a further modification of the metadata file.

In this paper we present SpLaSH (Spoken Language Search Hawk), a toolkit used for information retrieval from linguistic corpora. SpLaSH accepts data coming from different corpora and is able to merge annotations having different formats in the same database. Each annotation level uses its own set of temporal references, while no fixed hierarchy is imposed on the data model. SpLaSH provides tools to integrate time-aligned annotations (TMA) coded as Annotation Graphs [4], with text aligned ones (TXA), coded as generic XML files. 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 set of interfaces for querying the linguistic data.

The paper is organized as follows: Section 2 introduces the two annotation classes: TXA and TMA; Section 3 presents the SpLaSH data model; Section 4 shows how the system performs some queries.

2. Annotation classes

Spoken language corpora usually include both annotation describing acoustic properties of the speech signal, and processed texts resulting from speech transcriptions that were not necessarily made by listening to the speech signal. Therefore, in our model, analyses conducted on transcriptions are always time independent and produce labels aligned to sequences of linguistic units (usually words). Hence, our new system takes as input the two classes of linguistic annotation noted above: time-aligned (TMA) and text-aligned (TXA) annotation. For example, TMA labels are used to describe segmental phenomena (usually phones, syllables, morphemes, words...), and suprasegmental information about intonation, non verbal acoustic phenomena, disfluencies etc. TXA data usually contain morphosyntactic, semantic, pragmatic, and textual categories; annotators generally use various instances of XML markup to label their text.

2.1. TMA Annotations

TMA annotations are usually represented using various standard formats like TIMIT [5], PRAAT TextGrid “http://www.fon.hum.uva.nl/praat/”, PARTITUR [6], etc. The
lack of agreement on a common format for speech annotation means that similar phenomena can be labelled in quite different ways, making the annotations hard to compare.

Recently Bird and Liberman proposed that efforts should be focused on transforming existing datasets into a descriptive model compatible with the main annotation formats, thus defining a unifying standard to apply to each speech corpus. The solution proposed in this paper is based on an underlying logical structure that is common to all existing annotation formats. It comprises an oriented and acyclic graph (see figure 1) named Annotation Graph [4].

![Annotation Graph](figure1.png)

Figure 1: Annotation Graph logical and physical structure

The Annotation Graphs are data structures whose nodes are anchored to the signal and the annotations are represented by labels on the arcs. The graph is oriented with respect to the direction of the timeline. Each node contains timing information while the arcs are labeled with the type of annotation (like phones, words, syllables) and the value of annotation. Several relations are defined within arc- and node-like data structures. Temporal precedence relations are stored in the node fields while inclusion, coincidence and overlap relations are topologically expressed by the relative positions of arcs in the graph structure.

Our system implements TMA annotations by means of Annotation Graphs using the XML native database format originally proposed by Bird and Liberman. The formal definitions of the TMA data (based on AGs) and of TXA data (using generic XML format) are given in the following subsections:

### 3.1. Data Model

#### TMA Structure

Let C be a corpus. We define an XML tree according to the AG-XML data model. The main characteristics of SpLaSH are:

- The root of A has a label;
- Root child (primary) level is formed by a layer of gross segmentation (e.g., turns, breath groups etc.) of the examined text and an optional metadata section;
- At least one further layer of the tree is formed by nodes comprising a particular annotation level. The presence of further annotation levels introduces a hierarchy. Recursion is always allowed;
- Transcribed text (or a link referring to it) is represented at the level of the tree leaf.

### 3.2. TXA Annotations

TXA data can be recursive and often the descriptive elements which form the annotation system are hierarchically structured. XML is the ideal instrument for these types of annotations: it allows the sequential nature of the text, related to temporal development of speech units, to be inherent in the organization of the leaves. XML, being a metalanguage, ensures great freedom in the definition of the specific annotation system. To preserve this freedom, our system tries to maximally limit the number and nature of constraints that must be imposed on the TXA data in order to make those data coherent and connectible to the TMA. What follows is a formal description of the TXA data model definition:

**TXA Structure:** Let C be a corpus. We define an XML tree that respects the following conditions as a TexT-Aligned structure - TXA related to C:

- The root of A has a label;
- Root child (primary) level is formed by a layer of gross segmentation (e.g., turns, breath groups etc.) of the examined text and an optional metadata section;
- At least one further layer of the tree is formed by nodes comprising a particular annotation level. The presence of further annotation levels introduces a hierarchy. Recursion is always allowed;
- Transcribed text (or a link referring to it) is represented at the level of the tree leaf.

### 3.3. Spoken Language Search Hawk: SpLaSH

Spoken Language Search Hawk is a general purpose system for speech database management derived by annotated linguistic resources. The system integrates and processes linguistic data belonging both to TMA and TXA categories. The main characteristics of SpLaSH are:

- It presently converts TMA TIMIT and PRAAT-TextGrid annotation formats into AG-XML according to principles described in section 2.1;
- It processes any TXA XML format according to principles described in section 2.2;
- It offers tools for waveform visualization and listening: waveforms appear on the screen aligned to their TMA annotation;
- It offers tools for linguistic information retrieval.

### 3.4. Data Model

The main aim of SpLaSH is the integration of TMA and TXA data. As discussed in section 2, time aligned annotations are logically represented as graphs and text aligned ones as trees. Moreover both datasets are physically implemented in XML files. TXA Annotation values are stored in XML leaves while TMA ones are stored as arc attributes. We impose one simple constraint on these annotation classes in order to integrate them into an unique complex structure: TXA annotation values in the leaves must coincide with string units in at least one level of the TMA dataset.

Under this constraint, we introduce a new data structure named Connector Frame (CF) whose function is to act as an interface between the two original datasets. CF has the form of a tree with a root, a child level containing ID-values of TXA
nodes that are parents of textual leafs and finally a level containing ID-values of corresponding TMA values (see Figure 2). The integration between the TXA and TMA levels allows TXA to inherit the temporal relationships from the TMA levels.

The integration of linguistic data belonging to the TXA and TMA categories allows a user to perform sophisticated analysis on such data. The extraction of information from multi-level corpora is done by means of queries performed on both TMA and TXA structures. Queries can be performed in a Top-Down way (i.e. first on the TXA structure and second on the TMA structure), or in a Bottom-Up way (i.e. first TMA then TXA).

As illustrated in Figure 2, TMA annotations lie on an abstract horizontal plane X while the tree representing TXA annotations is on a vertical abstract plane Y, orthogonal to X. CF structure is represented by dotted lines.

4. Querying

The integration of linguistic data belonging to the TXA and TMA categories allows a user to perform sophisticated analysis on such data. The extraction of information from multi-level corpora is done by means of queries performed on both TMA and TXA structures. Queries can be performed in a Top-Down way (i.e. first on the TXA structure and second on the TMA structure), or in a Bottom-Up way (i.e. first TMA then TXA).

As illustrated in Figure 2, TMA annotations lie on an abstract horizontal plane X while the tree representing TXA annotations is on a vertical abstract plane Y, orthogonal to X. CF structure is represented by dotted lines.

Figure 2: Integrated TXA-TMA structure

As illustrated in Figure 2, TMA annotations lie on an abstract horizontal plane X while the tree representing TXA annotations is on a vertical abstract plane Y, orthogonal to X. CF structure is represented by dotted lines.

4.1. Simple Query

Simple queries can be performed on a TXA structure or on a TMA structure. Figure 4 shows the SpLaSH graphical interface used to perform queries on a single document TXA. Queries on TXA are strongly based on XPath operators (http://www.w3.org/TR/xpath20/). This interface is divided into several sections. The String XPath section is used to perform a query on TXA structures. The Builder Query section is composed of a set of dynamically generated graphic components used for the selection of nodes and attributes from the TXA corpus. The selection of elements within the components automatically generates an XPath expression that will be executed on the corpus. The Node result section displays the result of the XPath expression executed on the TXA corpus. If the XPath expression is generated dynamically in the Builder Query section, then the Nodes result section will show the intermediate results of the elements selection, being updated with every added element. Finally, the Translation of the query section shows the transformation in natural language of the XPath expression. This feature is useful for learning the meaning of XPath syntactic expressions.

SpLaSH presents an additional graphical interface for querying TMA corpora. In this interface there are several components used to express temporal relations between elements belonging to TMA annotations. Queries on TMA are strongly based on XQuery operators (http://www.w3.org/TR/xquery/).

4.2. Sequence Query

A sequence query is used to express the transitive closure over the arcs of annotation graphs. Using this type of query it is possible to select subsequences of contiguous annotations from a TMA corpus. The subsequence of contiguous annotations is specified by a sequence of strings called target sequence $T$. The target sequence can be composed of string values or of a combination of strings and the symbol ‘*’. The meaning assumed by the meta-character ‘*’ in a target sequence is $L^*$, where $L$ is the language used to write the linguistic annotation.

For example, the target sequence "word1 * word2" represents all the sequences of strings that begin with word1, end with word2 and contain any sequence of strings between them. So, specifying a target sequence $T$ and an annotation level $l_i$ in the TMA corpus, it is possible to find all contiguous annotation values of level $l_i$ belonging to the AGSet of the TMA corpus, which satisfies the target sequence $T$.

4.3. Cross Query

The cross queries interface, shown in Figure 5, is composed of two sections. The first section consists of a component used to express an XPath query on TXA structures. This component is similar to the String XPath section of the simple query interface for TXA structures. The second section is used for express-
a query that operates on the TMA data and presents a set of graphical components. These components allow: a) the selection of a TMA annotation level, Level; b) the constraints on the annotation value, Relation and Word search, and c) Relation type, which specifies the relationship that must exist between the TMA and TXA data. The cross queries are executed in the Top-Down direction and consist of a TXA node selection followed by a TMA arc selection. The TMA arc selection is performed on TMA data portions that underlie the nodes resulting from the TXA node selection. As shown in Figure 5, the results of the query are displayed in a table.

5. Conclusions

Compared with other corpora management systems presently available, SpLaSH presents interesting innovations and better performance.

SpLaSH does not impose a fixed hierarchy on annotation levels - other than those implicitly defined in the data model - as it is based on the idea that each level could be obtained, in principle, independently from the others. In this way different research groups can generate their labels with the minimal number of constraints necessary to integrate their work in a unified framework in further developments.

The next step is to enrich SpLaSH with a query language named SpLaSH-QL, whose formal definition is currently being finalized. SpLaSH-QL is formed by a set of specific algebraic operators aimed at the retrieval of information from the TXA and TMA integrated datasets. Operators can be composed and can explicitly contain XPath query recall as arguments. In the next SpLaSH releases, a guided interface for querying will be based on this language. Its use will increase the potential power of the query expression system.

SpLaSH is an open source project. A stable release of SpLaSH is freely available at http://s2snaples.fisica.unina.it/splash under GNU-Public license, and is constantly upgraded with new features.

6. References