Filtering multi-levels annotated data

Brigitte Bigi

Filtering multi-levels annotated data

Brigitte Bigi

Laboratoire Parole et Langage, CNRS, Aix-Marseille Univ.
5 avenue Pasteur, 13100 Aix-en-Provence, France
brigitte.bigi@lpl-aix.fr

Abstract

More and more annotated corpora are now available, and so are tools to annotate automatically and/or manually. As large multimodal-multiparty corpora become prevalent, new annotation and analysis requirements are emerging. Multimodal annotations are commonly organized in *tiers* each of which is a collection of *annotations*, each of which is commonly made of an anchor in time and a label. In this demo, we present DataFilter feature. It allows to define a set of filters to create new tiers with only the annotations matching the given filters. The system proposed in this demo is implemented as part of SPPAS software tool (Bigi, 2015), distributed under the terms of the GNU Public License.

1. Introduction

More and more annotated corpora are now available, and so are tools to annotate automatically and/or manually. As large multimodal-multiparty corpora become prevalent, new annotation and analysis requirements are emerging. Such multimodal-multiparty annotations are commonly organized in *tiers* each of which is a collection of *annotations*, each of which is commonly made of an anchor in time and a label.

"Corpora that include time-based data, such as video and marking gestures, make annotation and analysis of language and behavior much more complex than analysis based solely on text corpora and an audio signal" (Bigbee et al., 2001). Because annotating is not an end in itself, a minimum requirement after the annotation procedure is to automatically search for a subset of annotated entities and/or relationships between them: *linguists need to retrieve specific instances of a particular phenomenon.*

This demo presents a system for exploring and retrieving such multi-levels annotations. From a certain point of view, the purpose of the proposed system is **annotation mining**. From another point of view, the system could be considered like a data query but it doesn't need to convert annotated data into a database nor to write queries. Instead, the system proposes to combine **filters** which are applied directly on the given data and it retrieves the relevant instances in the same format as the given input ones so that they are organized in a new annotation layer.

2. Description of the system

In recent years, the SPPAS software tool (Bigi, 2015) has been developed by the author to automatically produce annotations and to analyze annotated data. SPPAS is multi-platform (Linux, MacOS and Windows) and open source issued under the terms of the GNU General Public License. It is specifically designed to be used directly by linguists.

This demo presents **DataFilter**, one of the features of SPPAS. A first version of this system was previously described in (Bigi and Saubesty, 2015). It has been already used in several studies like in (Tellier et al., 2012) to find correlations between speech and gestures, or in (Tellier et al., 2013) to find which gestures are produced while pausing, just to cite some of them.

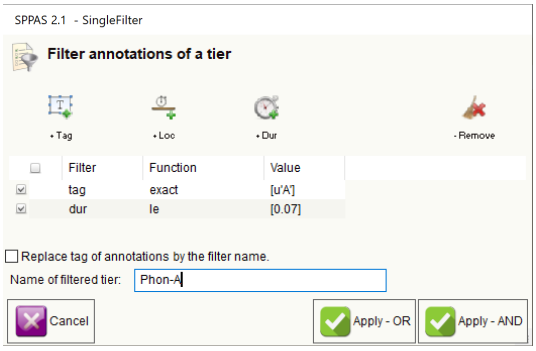

Figure 1: Example of filters with the GUI

DataFilter allows to define a set of filters to create new tiers made only of the annotations matching the given filters. Figure 1) illustrates the creation of filters with the Graphical User Interface (GUI). Two different types of filters can be created:

1. single filters to search for annotations in tiers depending on:
   - the label: an annotation is retrieved if its label exactly matches, contains, starts with or ends with the given pattern, or it is matching a regular expression. This filter can be case-sensitive or case-insensitive;
   - the anchor in time: an annotation is retrieved if it is located at a certain range of time in the timeline, i.e. before or after the given time value;
   - the duration: an annotation is retrieved if it's duration is greater, equal or lower than the given value.
2. relation filters to search for annotations of tiers that are in time-relation with annotations of another one. Time-relations are based on the Allen's interval algebra (Allen, 1983).

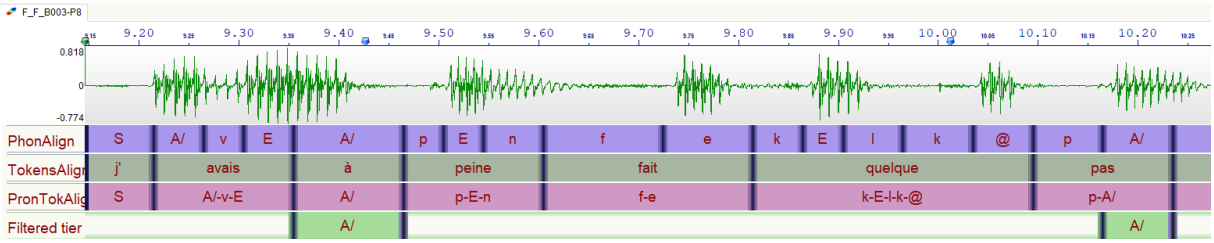

Figure 2: Example of filtered data

These filters can be combined with "`and`" and "`or`" operators and can be reverted.

From our point of view, this system results in the following advantages:

- It is expected to be powerful enough to meet the reasonable needs of end-users;
- It allows to import multi-levels annotated data from most of the existing annotation tools and to filter then save results directly in their native format (xra, TextGrid, eaf, ant, antx, csv, mrk, stm, ctm, ...);
- It provides the filtered annotated data in the form of a new annotation tier;
- It allows to filter multiple files at once;
- It can be used without requiring any XML-related or QL-related knowledge or skill;
- It proposes both a scripting language (Python 2.7 or Python 3.x) and a Graphical User Interface.

The proposed system was fully re-implemented recently for the following two main reasons. First, it's now doable *to add new filters* in the system. Secondly, it's easier *to write scripts* to filter data. Users of SPPAS observed only a few changes while using the GUI.

3. Example

DataFilter is illustrated by the next example writen with Python. It extracts annotations with `A/`[1] phonemes during less or equal than 70 ms from a tier with name 'PhonAlign':

```
tier = trs.find("PhonAlign")
filter = sppasFilter(tier)
phon_set = filter.tag(exact="A/") &
           filter.dur(le=0.07)
result = phon_set.to_tier("A-Phon")
```

Figure 1 illustrates a very close example, performed with the GUI. The checked lines are representing the filters. They were created by clicking on the button "+Tag" and filled the form then by clicking on "+Dur" and filling the form. The filters are applied when clicking on the button "Apply - AND".

[1] `A/` is the symbol in SAMPA encoding to represent both the sounds `A` and `a`

Figure 2 shows an extract of the result when the example is applied on the file F_F_B003-P8 which is one of the French samples freely available into the package of SPPAS. The phoneme `A/` of the 2nd token "avais" was not selected because its duration does not match the filter.

4. Conclusion & Future works

The system proposed in this demo is implemented as a feature of SPPAS: `http://www.sppas.org`. The documentation of such tool presents how to use this system in details for both the scripting language (section 6.5) and the GUI (section 5.6).

Future developments will focus on creating new filters, for example to retrieve annotations depending on their occurrences: less or more than a given value.

5. References

Allen, James-F., 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*:832–843.

Bigbee, Tony, Dan Loehr, and Lisa Harper, 2001. Emerging requirements for multi-modal annotation and analysis tools. In *INTERSPEECH*. Aalborg, Denmark.

Bigi, Brigitte, 2015. SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, 111–112:54–69.

Bigi, Brigitte and Jorane Saubesty, 2015. Searching and retrieving multi-levels annotated data. In *Proceedings of Gesture and Speech in Interactioni - 4th edition*. Nantes, France.

Tellier, Marion, Gale Stam, and Brigitte Bigi, 2012. Same speech, different gestures? In *5th International Society for Gesture Studies*. Lund, Sweden.

Tellier, Marion, Gale Stam, and Brigitte Bigi, 2013. Gesturing while pausing in conversation: Self-oriented or partner-oriented? In *The combined meeting of the 10th International Gesture Workshop and the 3rd Gesture and Speech in Interaction conference*. Tillburg, The Netherlands.