SPPAS - Multi-Lingual approaches to the automatic annotation of speech

Brigitte Bigi

September, 29th, 2016

1

Presentation

1.1 Corpus and annotation

- Corpus linguistics is the study of language as expressed in samples (corpora) of "real world".
- Corpus annotation is a path to greater linguistic understanding.

> Corpus annotation "can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process" (Leech, 1997).

1.2 Annotations

- Annotations must be time-synchronized:
  - annotations need to be time-aligned in order to be useful for purposes such as qualitative or quantitative analyses
- Temporal information makes it possible to describe simultaneous behaviours:
  - of different levels in an utterance (e.g. prosody and locution);
  - of different modalities (e.g. speech and gesture);
  - of different speakers or extralinguistic events.
- Time-analysis of multi-level annotations can reveal linguistic structures

1.3 Time-synchronized annotations

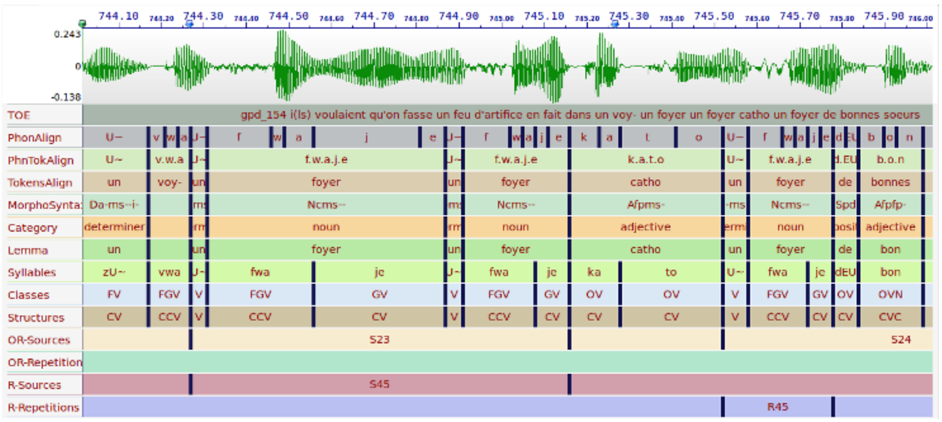

744.10 744.20 744.30 744.40 744.50 744.60 744.70 744.80 744.90 745.00 745.10 745.20 745.30 745.40 745.50 745.60 745.70 745.80 745.90 746.00
0.243
0
-0.138
TOE gpd_154 i(ls) voulaient qu'on fasse un feu d'artifice en fait dans un voy- un foyer un foyer catho un foyer de bonnes soeurs
PhonAlign U~ v w a U~ f w a j e U~ f w a j e k a t o U~ f w a j e d EU b o n
PhnTokAlign U~ v.w.a U~ f.w.a.j.e U~ f.w.a.j.e k.a.t.o U~ f.w.a.j.e d.EU b.o.n
TokensAlign un voy- un foyer un foyer catho un foyer de bonnes
MorphoSynta Da-ms--i- ms Ncms-- ms Ncms-- Afpms- ms Ncms-- Spd Afpfp-
Category determiner rm noun rm noun adjective erm noun posit adjective
Lemma un un foyer un foyer catho un foyer de bon
Syllables zU~ vwa U~ fwa je U~ fwa je ka to U~ fwa je dEU bon
Classes FV FGV V FGV GV V FGV GV OV OV V FGV GV OV OVN
Structures CV CCV V CCV CV V CCV CV CV CV V CCV CV CV CVC
OR-Sources S23 S24
OR-Repetition
R-Sources S45
R-Repetitions R45

Figure 1.1: Example of multi-level annotations

1.4 Annotation software

- Manual annotation
  - Audacity;
  - Praat;
  - Elan;
  - Anvil;
  - Winpitch;
  - . . .
- Automatic annotation
  - The current state-of-the-art in Computational Linguistics allows many annotation tasks to be semi- or fully- automated.
- Interoperability: when such muti-layer corpora are to be created with existing task-specific annotation tools, a new problem arises: output formats of the annotation tools can differ considerably.

1.5 Automatic annotation

- Annotation is not an end in itself - it is a basis for further analysis.
- Each annotation that *can* be done automatically *must* be done automatically!
- Why? Because *revising* is faster and easier than *annotating*. . . if the automatic system is "good enough".

Before using any automatic annotation tool/software, it is important to consider its error rate (where applicable) and to estimate how those errors will affect the purpose for the annotated corpora.

1.6 SPPAS: the automatic annotation and analysis of speech

- SPPAS is a multi-platform and public annotation software tool:
  - It is able to produce automatically speech annotations from a recorded speech sound and its orthographic transcription.
  - Some special features are also offered for the analysis of any kind of annotated files.
- SPPAS is compatible with Praat, Elan, Transcriber, Annotation Pro, Phonedit, and many others...

1.7 Automatic annotations: main advantage

- Language-independent algorithms implie:
  - language-dependent resources;
  - easy and fast to add a new language;
  - easy and fast to modify existing resources.

1.8 Summary

- **Phonemes and words segmentation**
- Syllables segmentation
- Repetitions detection
- Conclusion

2

Phonemes and words segmentation

2.1 Definition

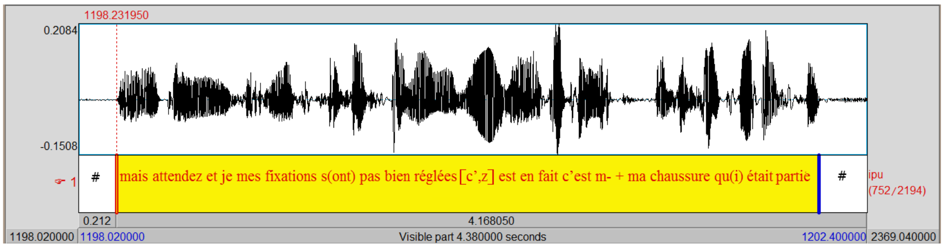
- the process of taking the orthographic transcription text of an audio speech segment, like *IPUs*, and determining where particular phonemes/words occur in this speech segment.

IPUs = Inter-Pausal Units

2.2 Data preparation

- Audio file with the following recommended conditions:
  - one file = one speaker
  - good recording quality (anechoic chamber)
  - 16000Hz, 16bits
- Orthographic transcription:
  - follow the convention of the software
  - enriched with: filled pauses; short pauses; truncated words; repeats; noises and laugh items.

2.2.1 Data preparation: example

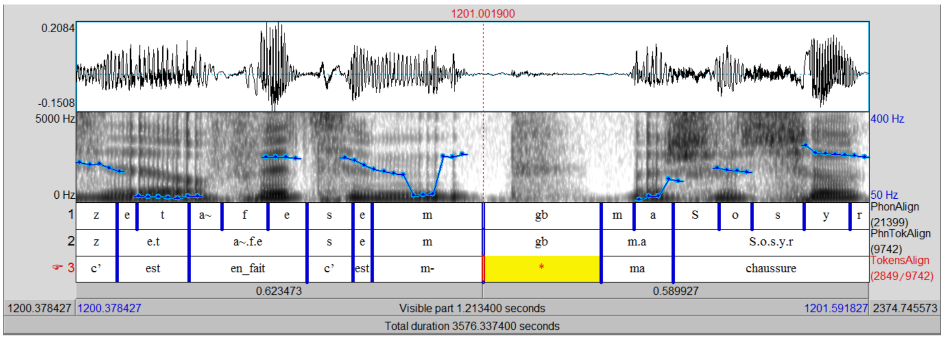


mais attendez et je mes fixations s(ont) pas bien réglées [c',z] est en fait c'est m- + ma chaussure qu(i) était partie

Figure 2.1: An IPU of "Corpus of Interactional Data"

2.3 Expected result

- Time-aligned phonemes and tokens and events like noises or laughter



2.4 Phonemes and words segmentation in SPPAS

1. tokenization
   - text normalization
2. phonetization
   - grapheme to phoneme conversion
3. alignment
   - speech segmentation

All three tasks are fully-automatic, but each annotation output can be manually checked if desired.

2.4.1 Tokenization

- Tokenization is also known as "Text Normalization".
- Tokenization is the process of segmenting a text into tokens.
- In principle, any system that deals with unrestricted text need the text to be normalized.
- Automatic text normalization is mostly dedicated to written text, in the NLP community

> Text normalization development is commonly carried out specifically for each language and/or task even if this work is laborious and time consuming. Actually, for many languages there has not been any concerted effort directed towards text normalization.

2.4.2 Tokenization in SPPAS

- SPPAS implements a generic approach:
  - a text normalization method as language and task independent as possible.
  - This enables adding new languages quickly when compared to the development of such tools from scratch.
- This method is implemented as a set of modules that are applied sequentially to the text corpora.
- The portability to a new language consists of:
  - inheriting all language independent modules;
  - (rapid) adaptation of other language dependent modules.

2.4.3 Tokenization main steps

- Remove punctuation
- Lower the text
- Convert numbers to their written form
- Replace symbols by their written form (like %, °, . . .):
  - based on a lexicon.
- Word segmentation:
  - based on a lexicon.

2.4.4 Tokenization of speech transcription

- two types of transcriptions are automatically derived by the automatic tokenizer:
  1. the “standard transcription” (a list of orthographic tokens/words);
  2. the “faked transcription” that is a specific transcription from which the obtained phonetic tokens are used by the phonetization system.

2.4.5 Tokenization: example

- Transcription:
  - mais attendez et je mes fixations s(ont) pas bien réglées [c’,z] est en fait c’est m- + ma chaussure qu(i) était partie
- Standard tokenization:
  - mais attendez et je mes fixations sont pas bien réglées c’ est en_fait c’est ma chaussure qui était partie
- Faked tokenization:
  - mais attendez et je mes fixations s pas bien réglées z est en_fait c’est m- + ma chaussure qu était partie

---

2.4.6 Tokenization: reference

```
Brigitte Bigi (2014).
A Multilingual Text Normalization Approach.
Human Language Technologies Challenges for Computer Science and Linguistics.
LNAI 8387, Springer, Heidelberg. ISBN: 978-3-319-14120-6. Pages 515-526.
```

2.5 Phonetization

- Phonetization is also known as grapheme-phoneme conversion
- Phonetization is the process of representing sounds with phonetic signs.
- Phonetic transcription of text is an indispensable component of text-to-speech (TTS) systems and is used in acoustic modeling for automatic speech recognition (ASR) and other natural language processing applications.

  Converting from written text into actual sounds, for any language, cause several problems that have their origins in the relative lack of correspondence between the spelling of the lexical items and their sound contents.

2.5.1 Phonetization in SPPAS

- SPPAS implements a generic approach:
  - consists in storing a maximum of phonological knowledge in a lexicon.
  - In this sense, this approach is language-independent.
- The phonetization process is the equivalent of a sequence of dictionary look-ups.
- An important step is to build the pronunciation dictionary, where each word in the vocabulary is expanded into its constituent phones, including pronunciation variants.

2.5.2 Phonetization of normalized speech transcription

- SPPAS implements a language-independent algorithm to phonetize unknown words
  - given enough examples (in the dictionary) it should be possible to predict the pronunciation of unseen words purely by analogy.
- Example with unknown word “pac-aix”:
  - English: p-{-k-aI-k-s|p-{-k-eI-aI-k-s|p-{-k-aI-E-k-s|p-{-k-eI-aI-E-k-s
  - French: p-a-k-E-k-s
  - Mandarin Chinese: p_h-a-a-i

2.5.3 Phonetization: example

- Tokenization:
  - mais attendez je
- Phonetization:
  - m-E-z|m-e|m-E|m-e-z|m a-t-a~-d-e|a-t-a~-d-e-z e|E Z|Z-eu|S

2.5.4 Phonetization: reference

```
Brigitte Bigi (2016).
A phonetization approach for the forced-alignment task in SPPAS.
Human Language Technologies Challenges for Computer Science and Linguistics.
LNAI 9561, Springer, Heidelberg.
```

2.6 Alignment

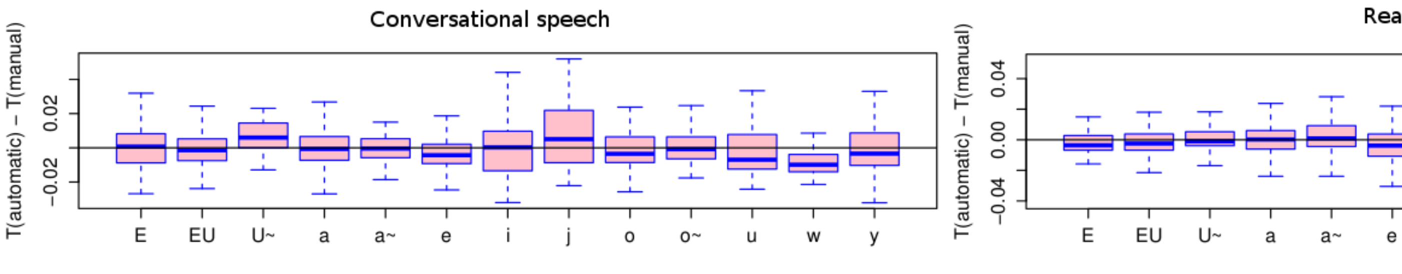
- Alignment is also called phonetic segmentation
- The alignment problem consists in a time-matching between a given speech unit along with a phonetic representation of the unit.
- Many freely available tool boxes, i.e. Speech Recognition Engines that can perform Speech Segmentation
  - HTK - Hidden Markov Model Toolkit
  - CMU Sphinx
  - Open Source Large Vocabulary CSR Engine Julius
  - . . .

2.7 Alignment in SPPAS

- SPPAS is based on Julius.
- This choice is motivated by four main reasons:
  1. the Julius toolkit is open-source, so there is no specific reason to develop a new one;
  2. it is easy to install which is important for end-users;
  3. it's usage is relatively easy so it was convenient to integrate it in SPPAS;
  4. its performance corresponds to the state-of-the-art of other available systems of such kind.

2.7.1 Alignment: results in SPPAS

- In average, automatic speech segmentation of French is 95% of times within 40ms compared to the manual segmentation (SPPAS 1.5, September 2014)



2.7.2 Alignment: references

```
Brigitte Bigi (2012).
The SPPAS participation to the Forced-Alignment task of Evalita 2011.
B. Magnini et al. (Eds.): EVALITA 2012, LNAI 7689, pp. 312-321. Springer, Heidelberg.

Brigitte Bigi (2014).
The SPPAS participation to Evalita 2014.
In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014
and the Fourth International Workshop EVALITA 2014, Pisa, Italy.

Brigitte Bigi (2014).
Automatic Speech Segmentation of French: Corpus Adaptation.
In 2nd Asian Pacific Corpus Linguistics Conference, pp. 32, Hong Kong.
```

2.8 Summary

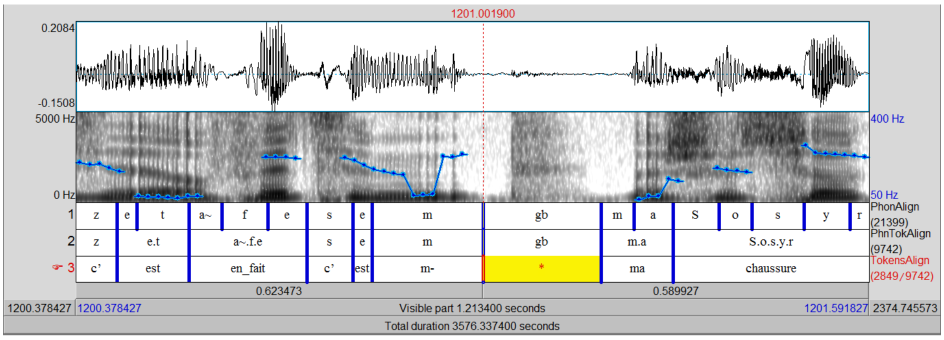
- Phonemes and words segmentation
- **Syllables segmentation**
- Repetitions detection
- Conclusion

3

Syllables segmentation

3.0.1 Data preparation

- Time-aligned phonemes

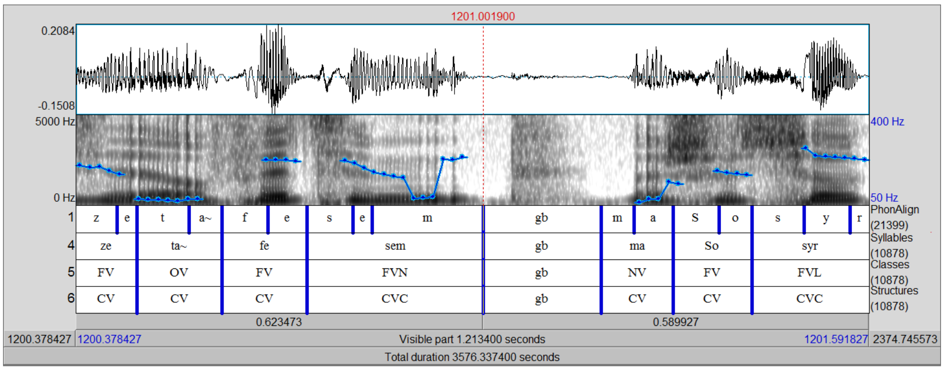

1201.001900
0.2084
-0.1508
5000 Hz
0 Hz
400 Hz
50 Hz
1 z e t a~ f e s e m gb m a S o s y r
2 z e.t a~.f.e s e m gb m.a S.o.s.y.r
3 c' est en_fait c' est m- * ma chaussure
PhonAlign (21399)
PhnTokAlign (9742)
TokensAlign (2849/9742)
0.623473
0.589927
1200.378427
1200.378427
Visible part 1.213400 seconds
1201.591827
2374.745573
Total duration 3576.337400 seconds

3.1 Expected result

- Time-aligned syllables

3.1.1 Syllabification in SPPAS

- A rule-based system to cluster phonemes into syllables
- Rules available for:


1201.001900
0.2084
-0.1508
5000 Hz
0 Hz
400 Hz
50 Hz
1 z e t a~ f e s e m gb m a S o s y r
4 ze ta~ fe sem gb ma So syr
5 FV OV FV FVN gb NV FV FVL
6 CV CV CV CVC gb CV CV CVC
PhonAlign (21399)
Syllables (10878)
Classes (10878)
Structures (10878)
0.623473
0.589927
1200.378427
1200.378427
Visible part 1.213400 seconds
1201.591827
2374.745573
Total duration 3576.337400 seconds

- French
- Italian
- Polish

- This phoneme-to-syllable segmentation system is based on 2 main principles:

  1. a syllable contains a vowel, and only one;
  2. a pause is a syllable boundary.

3.1.2 Syllabification in SPPAS

- Phonemes are grouped into classes
- Classes for both French and Italian:
  - V - Vowels,
  - G - Glides,
  - L - Liquids,
  - O - Occlusives,
  - F - Fricatives,
  - N - Nasals.
- Fix rules to find the boundaries between two vowels

3.1.3 Syllabification: references

Brigitte Bigi, Christine Meunier, Irina Nesterenko and Roxane Bertrand (2010).
Automatic detection of syllable boundaries in spontaneous speech.
Language Resource and Evaluation Conference, pages 3285-3292, La Valetta, Malte.

Brigitte Bigi and Caterina Petrone (2014).
A generic tool for the automatic syllabification of Italian.
In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and
of the Fourth International Workshop EVALITA 2014, pp. 73-77, Pisa, Italy.

Brigitte Bigi and Katarzyna Klessa (2015).

Automatic Syllabification of Polish.
In 7th Language and Technology Conference: Human Language Technologies as a Challenge for
Computer Science and Linguistics, pp. 262-266, Poznan, Poland.

3.2 Summary

- Phonemes and words segmentation
- Syllables segmentation
- **Repetitions detection**
- Conclusion

4

Repetitions detection

4.1 Repetitions

- Other-repetition is a device involving the reproduction by a speaker of what another speaker has just said.
- Other-repetition has been identified as an important mechanism in face-to-face conversation through their discursive or communicative functions.
- Example:

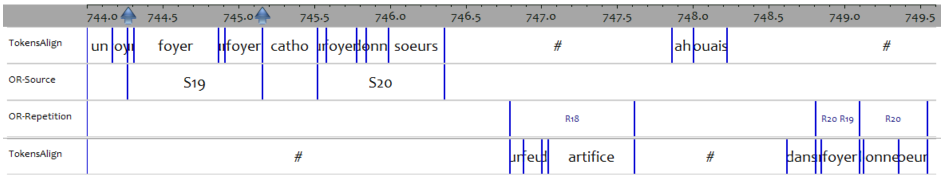
Figure 4.1: Extract of Corpus of Interactional Data

```
AB: ils voulaient qu'on fasse un feu d'artifice en fait dans un voy- un foyer un foyer catho
un foyer de bonnes soeurs

CM: un feu d'artifice

AB: ah ouais

CM: dans un foyer de bonnes soeurs

CM: @
```

4.1.1 Repetitions in SPPAS

- SPPAS implements:
  - *self-repetitions*;
  - *other-repetitions* detection.
- The system is based only on lexical criteria
  - observable cues: time-aligned tokens converted to lemmas (if any).
- The system was used to propose a lexical characterization of OR: various statistics was estimated on the detected OR.

4.1.2 Repetitions: other-repetitions detection result



4.1.3 Repetitions: reference

```
Brigitte Bigi, Roxane Bertrand, Mathilde Guardiola (2014).
Automatic Detection of Other-Repetition Occurrences: Application to French Conversational Speech
In Proceedings of the Ninth International Conference on Language Resources and Evaluation,
pp. 836-842, Reykjavik, Iceland.
```

4.2 Summary

- Phonemes and words segmentation
- Syllables segmentation
- Repetitions detection
- **Conclusion**

5

Conclusion

5.1 SPPAS: multi-lingual approaches

- SPPAS is a computer software tool designed and developed to handle multiple language corpora and/or tasks with the same algorithms in the same software environment.
- SPPAS emphasizes new practices in the methodology of tool developments:
  - considering problems with a generic multi-lingual aspect,
  - sharing resources,
  - putting the end-users in control of their own computing.
- Only the resources are language-specific and the approach is based on the simplest resources possible.

5.2 Resources extend

- Phoneticians are of crucial importance for resource development
  - they can contribute to improve the resources used by automatic systems.
- New versions are systematically released to the public and serve to benefit of the whole community.
- Resources are distributed under the terms of a public license, so that SPPAS users:
  - have free access to the application source code and the resources of the software they use,
  - are free to share the software and resources with other people,
  - are free to modify the software and resources,
  - are free to publish their modified versions of the software and resources.

5.3 About SPPAS

```
Brigitte Bigi (2015).
SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech.
The Phonetician, 111-112, pp. 54-69.
```

- Web site:

  http://www.sppas.org