SPPAS tutorial: Methodology and software for the semi-automatic annotation and analysis of speech

Brigitte Bigi

September, 28th, 2015

1

Presentation

1.1 Summary

- Introduction (25')
- Selection of annotation software (15')
- Corpus development methodology (45')
- SPPAS (30')
- Conclusion (5')

1.2 Presenter: Brigitte Bigi

- Researcher at the **CNRS**, Laboratoire Parole et Langage, Aix-Marseille Université, Aix-en-Provence, France

Figure 1.1: Brigitte Bigi

- Computer Scientist working in the field of corpora and annotations:
  - formalization/constitution of corpora,
  - automatic annotation (mainly at the phonetic level, also at the discourse level),
  - multimodality (annotation, exploration, extraction of annotated data),
  - multilinguality (methods and algorithms).
- Author and developer of SPPAS - Automatic Annotation of Speech

---

1.3 Tutorial scopes

- This tutorial will report on methodology for the manual and/or automatic annotation and analysis of a recorded speech corpus.
- We illustrate the steps to take in the perspective of:
  - obtaining rich and broad-coverage speech annotation
  - and initial analysis of such a corpus
  - both with a specific focus on SPPAS software.

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).

1.4 Summary

- **Introduction**
- Selection of annotation software
- Corpus development methodology
- SPPAS
- Conclusion

2

Introduction

2.1 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 and rigour:
  - The annotation of recordings is practised by many Linguistics sub-fields, such as Phonetics, Prosody, Gesture or Discourse...
  - Corpora are annotated with detailed information at various linguistic levels thanks to **annotation software(s)**.
  - New requirements are emerging for **very large multimodal corpora** where manual analysis is impractical.

2.1.1 Multi-domain 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
- Annotation requires software

2.2 Annotation software

- Manual annotation
- Automatic annotation
  - The current state-of-the-art in Computational Linguistics allows many annotation tasks to be semi- or fully- automated.
- But...
  1. Despite these advances that have been achived for annotating and analysing language, many annotation frameworks and/or models for the construction and analysis of multimodal data continue to rely on "low-tech" and/or manual technologies.
  2. 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.

2.3 A methodology for annotation...

- Annotation is not an end in itself - it is a basis for further analysis
- Handling of 'Big data', consisting of large quantities of audio, audio-visual and other multimodal recordings, is beyond the capabilities of purely manual annotation and traditional manual statistical analysis and plotting
- Two phases of automation are needed:
  - The Automatic Annotator
  - The Automatic Analyzer

2.3.1 Corpus annotation: Manual vs. Automatic

- The wide range of annotations, from aligned transcripts to gaze to reference to gestural form, is costly to collect and to annotate, both in terms of time and money.
- 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".

2.3.2 The Automatic Annotator

- The Automatic Annotator time-aligns descriptive data for Tiers such as Phonetics, Prosody, Syntax, Discourse with the recorded signal:

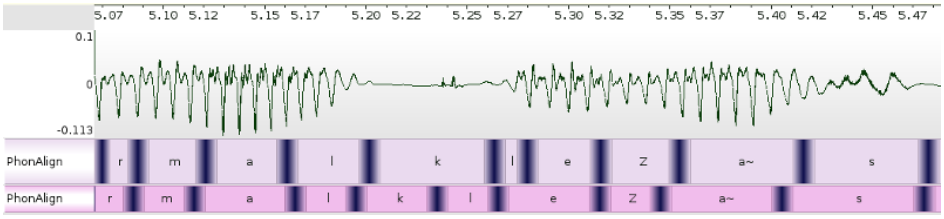


Figure 2.1: Example of automatic time-alignment vs manual time-alignment

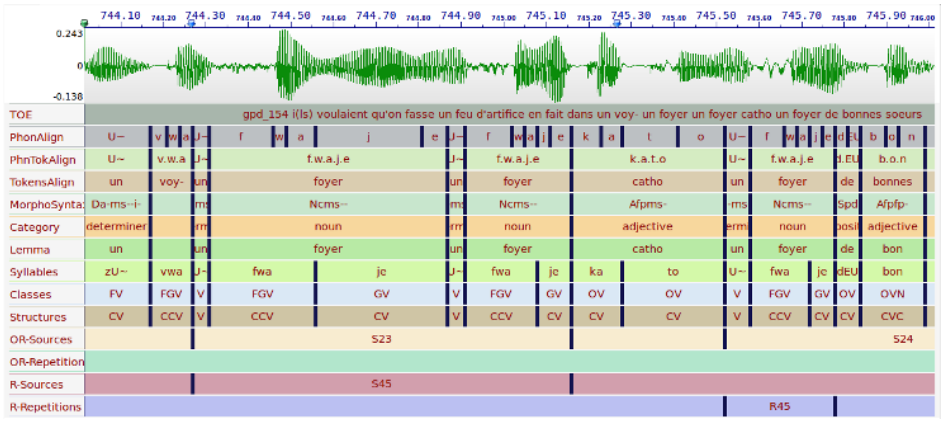


Figure 2.2: Example of multi-level annotations: only the orthographic transcription is manual

2.3.3 The Automatic Analyzer

- The output of the Automatic Annotator is usually manually post-edited before being input to the Automatic Analyzer
- The Automatic Analyzer inputs time-aligned data and outputs a report
  - about annotation Labels
  - sequences of annotation Labels in annotation Tiers
  - relations between Labels in sets of annotation Tiers
  - with statistics
  - with visualisations

2.4 Getting/Sharing a corpus

- Maybe there is already a corpus you can use?
- Data reporitories: depending on the research discipline, data can often be deposited in one or more data centers (or repositories) that will provide access to the data. These repositories may have specific requirements:
  - subject/research domain
  - data re-use and access
  - file format and data structure, and
  - metadata.

- SLDR:
  - http://sldr.org
  - Speech and Language Data Repository
  - gathering and sharing language data
  - long-term preservation by CINES, an institutional archive site.

2.5 Corpora

- CID - Corpus of Interactional Data
- GrenelleII corpus:
  - http://sldr.org/sldr000744
- Aix MapTask:
  - http://sldr.org/sldr000732
  - http://sldr.org/sldr000875
- DVD corpus:
  - http://sldr.org/sldr000891

Figure 2.3: Screenshots of 4 corpora (left to right): CID, GrenelleII, Aix MapTask, DVD

2.5.1 CID - Corpus of Conversational Data

- Face-to-face conversations in French
- Created by Roxane Bertrand and Béatrice Priego-Valverde
- 8 semi-guided dialogs (110,000 words)
- Recorded in 2003 and 2005
- Available at:
  - http://sldr.org/sldr000027/
  - http://sldr.org/sldr000720/
- Corpus description: `(Bertrand et al. 2008)`
- Multimodal annotations: `(Blache et al. 2010)`

2.5.2 CID - Extracts

2.5.3 CID - a pioneer

- No annotation framework nor tools were available
- Two many data to manually annotate at all levels!

Then...

1. an annotation scheme was developed for each annotation level
2. the framework I'm currently presenting was elaborated
3. automatic tools were adapted or designed
4. a multi-level request system was designed

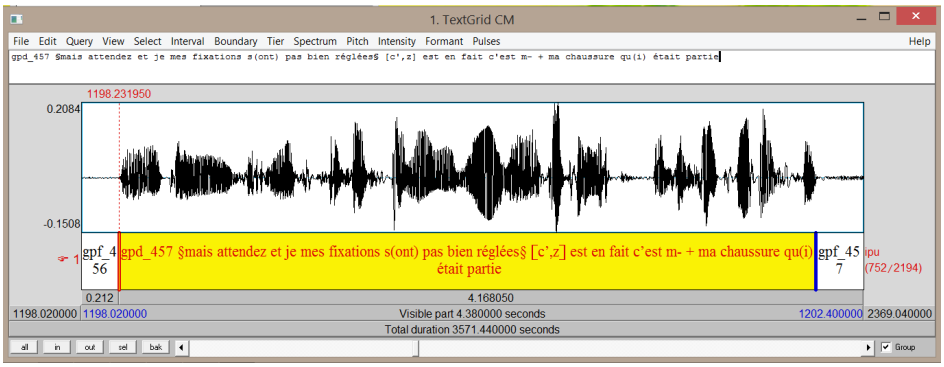
... annotated either by LPL, LLING or LIMSI.

2.5.4 CID - Current annotations (1)

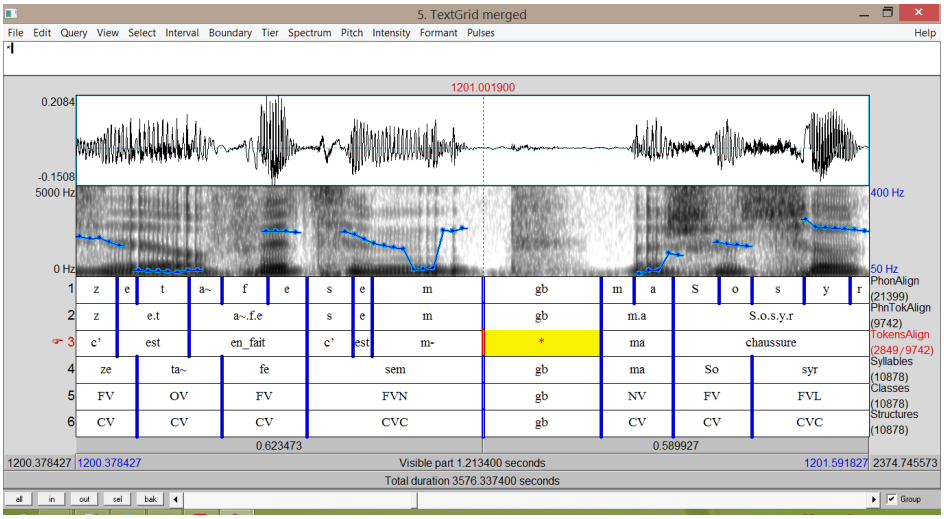
1. Enriched orthographic transcription (manual)
    - time-aligned at the IPU level (automatic)

2.5.5 CID - Current annotations (2)

2. Time-aligned phonemes and tokens and events like noises, laughter (automatic)
3. Time-aligned syllables (automatic)

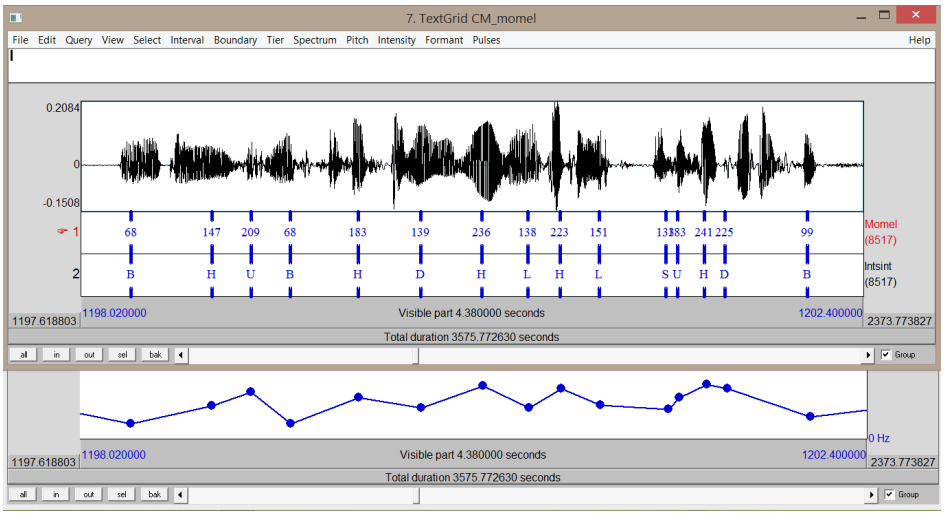

1. TextGrid CM
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
gpd_457 §mais attendez et je mes fixations s(ont) pas bien réglées§ [c',z] est en fait c'est m- + ma chaussure qu(i) était partie
1198.231950
0.2084
-0.1508
gpf_456
gpd_457 §mais attendez et je mes fixations s(ont) pas bien réglées§ [c',z] est en fait c'est m- + ma chaussure qu(i) était partie
gpf_457
0.212
4.168050
1198.020000 1198.020000 Visible part 4.380000 seconds 1202.400000 2369.040000
Total duration 3571.440000 seconds


0:00
0:0


5. TextGrid merged
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
1201.001900
0.2084
-0.1508
5000 Hz
0 Hz
400 Hz
50 Hz
PhonAlign (21399)
PhnTokAlign (9742)
TokensAlign (2849,/9742)
Syllables (10878)
Classes (10878)
Structures (10878)
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
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
0.623473 0.589927
1200.378427 1200.378427 Visible part 1.213400 seconds 1201.591827 2374.745573
Total duration 3576.337400 seconds

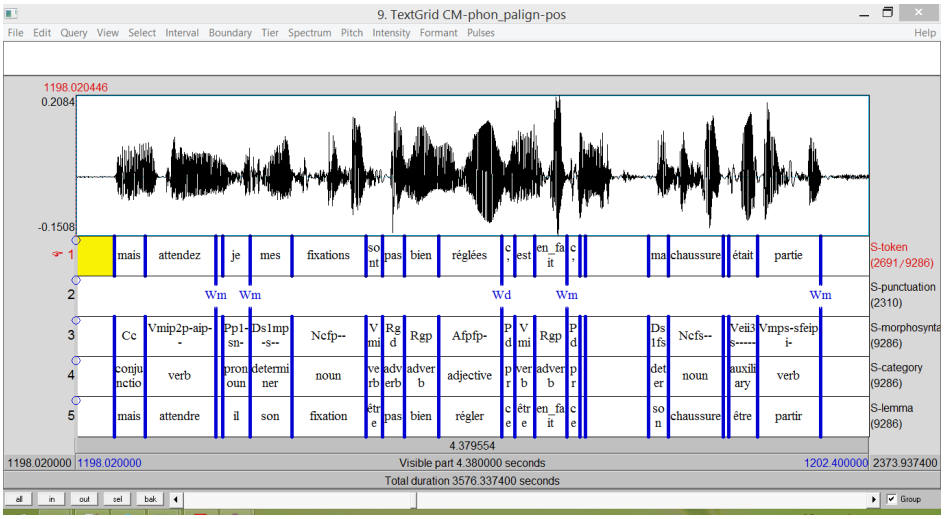
2.5.6 CID - Current annotations (3)

4. Prosodic contours (manual)
5. Momel - Modelization of melody (automatic)
6. **IN**ternational **T**ranscription **S**ystem for **INT**onation (automatic)



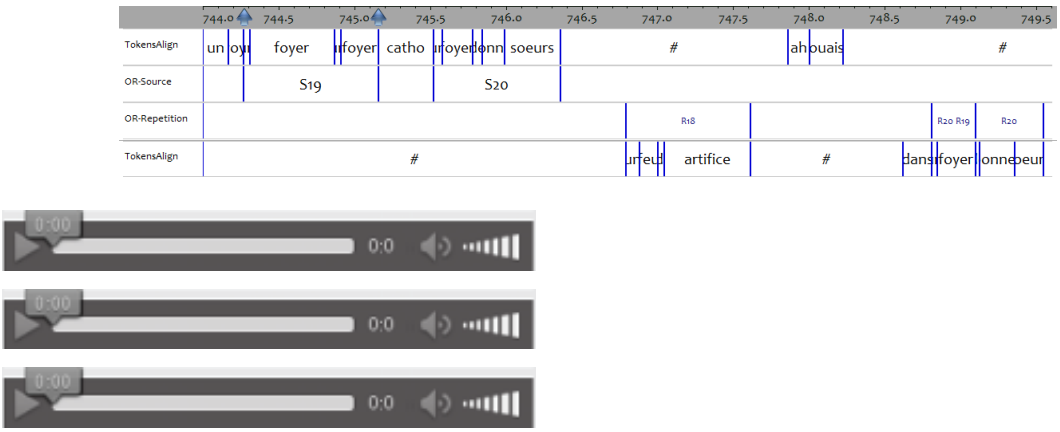
2.5.7 CID - Current annotations (4)

7. Morpho-syntax and syntax time-aligned at the token level (automatic);
8. Time-aligned lemmas (automatic);



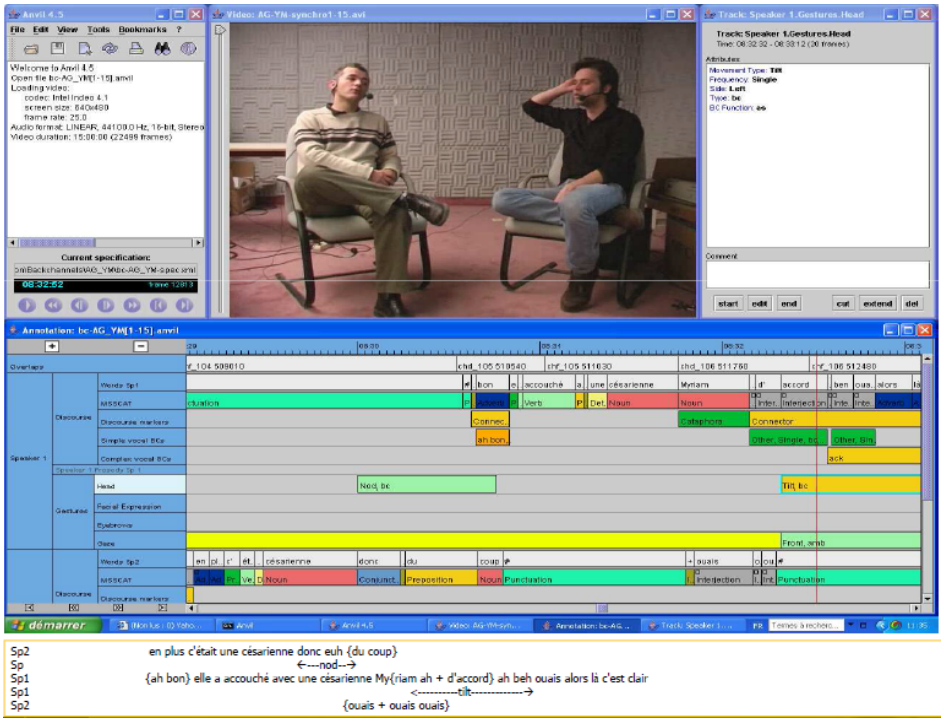
2.5.8 CID - Current annotations (5)

9. Dysfluencies (manual)
10. Discourse and interaction (manual)
11. Other- and Self- Repetitions (semi-automatic)



2.5.9 CID - Current annotations (6)

12. Gestures: postural, face, hands (manual)



2.5.10 CID - to summarize

- 8 face-to-face conversations
- A very (very very) large number of time-aligned annotations
- An annotation methodology and annotation tools/software
- More than 80 publications in 2013

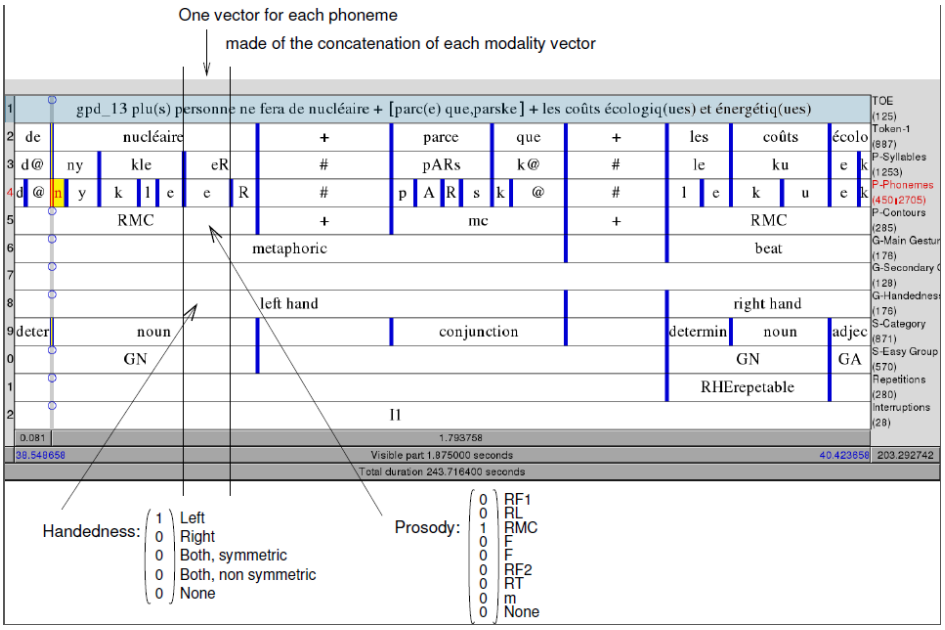
2.5.11 GrenelleII

- Video downloaded from a FTP server (after authorization), a flv file with poor quality
- Audio extracted from the video

2.5.12 GrenelleII: annotations

1. Enriched orthographic transcription (manual)
    - time-aligned at the utterance level (automatic)
2. Time-aligned phonemes, tokens and events (automatic)
3. Time-aligned syllables (automatic)
4. Prosodic contours and intonation (manual)
5. Morpho-syntax time-aligned at the token level (automatic)
6. Self-repetitions (semi-automatic)
7. Interruptions (manual)

2.5.13 GrenelleII: Multi-modal analysis



2.5.14 Aix Map-Task

- A French Map-Task
- Available at:
  - http://sldr.org/sldr000732
  - http://sldr.org/sldr000875
- 8 maps for each pair of speakers
- 2 recording sessions:
  - 2002: Remote condition, 4 dialogs, audio
  - 2013: Face-to-face condition, 5 dialogs, audio + video
  - the same maps for both sessions
- (Bard et al. 2013), (Gorish et al. 2014)

2.5.15 Aix Map-Task: Screenshot

2.5.16 Aix Map-Task: Annotations

1. Enriched orthographic transcription (manual)
   - time-aligned at the utterance level (manual in 2002 / automatic in 2013)
2. Time-aligned phonemes and tokens and events (automatic)
3. Time-aligned syllables (automatic)
4. Feedback (semi-automatic)

Figure 2.4: Face to face Aix Map-Task

2.6 Why a rigorous methodology?

- Quick and dirty annotation is possible, unless you expect to:
  1. produce reliable annotations
  2. perform complex analysis
  3. re-use annotations
  4. share the corpus and its annotations

Figure 2.5: Quick and dirty prototyping

2.7 Summary

- Introduction
- **Selection of annotation software**
- Corpus development methodology
- SPPAS
- Conclusion

3

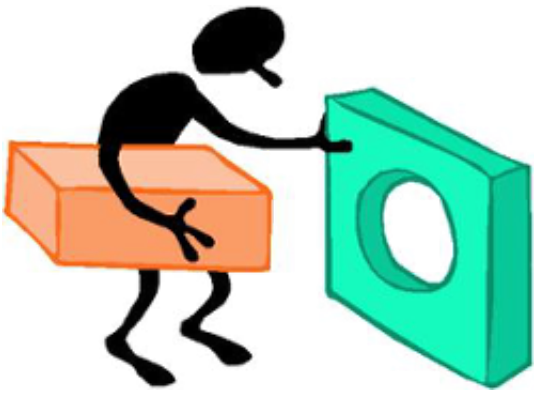
Selection of annotation software

3.1 Introduction

- In recent years, many annotation software (or tools) have become available for annotation of digital audio-video data.
- For a researcher looking for an annotation software, it is difficult to decide about its usefulness and usability.
- Some are mainly dedicated to Computer Scientists (tools) and some are designed for Linguists (software), some are designed for both.

3.2 Software selection

- The choise of all annotation software must be done carefully, and **before** the creation of the corpus.
- It is part of the corpus creation framework.


versus


3.2.1 Software selection: requirements

To decide about usefulness and usability, it is necessary to know all of the followings:

1. about the license,
2. about the ease of use,
3. about the strengths/weaknesses for specific annotation purposes,
4. about the type of data or analysis the tool/software is designed for,
5. about its compatibility with other annotated data.

3.2.2 Finding and evaluating appropriate software (1)

1. About the license

- prefer free and open source software:

Even if you can personally afford to pay for a licence for software you may wish to share your methodology with other students or researchers who cannot afford to buy a license.

3.2.3 Finding and evaluating appropriate software (2)

2. About the ease of use

- software or web-services?
- prefer multi-platform software:
  - different scientific communities tend to use Mac OS, Windows or Unix platforms.
  - multi-platform software makes sharing between such communities much easier.

- GUI or CLI usability: prefer usable software!

  If the software requires the help of an engineer each time you need to use it, this will be a serious limitation on your usage.

3.2.4 Finding and evaluating appropriate software (3)

3. About the strengths/weaknesses for specific annotation purposes

- Investigate whether the software has been found to be reliable and is likely to improve the efficiency of workflow, and either accelerate your work or enable you to deal with more extensive data, or both.

3.2.5 Finding and evaluating appropriate software (4)

4. About the type of data or analysis the tool/software is designed for

When annotating corpora at multiple linguistic levels, annotators may use different expert tools for different phenomena or types of annotation. These tools employ different data models and accompanying approaches to visualization, and they produce different output formats. (Chiarcos et al. 2008)

3.2.6 Finding and evaluating appropriate software (5)

5. About its compatibility with other annotated data

- None of the software are interoperable (open/save files)
- Prefer compatible software (import/export files)

- Estimate the availability to import/export data with a minimum loss of information

---

3.3 Automatic vs. Manual

- Manual:
  - Linguistics data are annotated several times by one or several annotators, each one annotates according to his/her knowledge, beliefs and uncertainty.
- Automatic annotation tools/software:
  - add detailed information to language data on the basis of procedures written into the software, without human intervention other than to run the program.
  - sometimes performed by following rules set out by programmers and linguists,
  - most often, annotation programs are at least partly based on machine learning algorithms that are trained using manually annotated examples.

3.3.1 Automatic vs. Manual

> Highly reliable manually annotated resources are, naturally, more expensive to construct, rarer and smaller in size than automatically annotated data, but they are essential for the development of automated annotation tools and are necessary whenever the desired annotation procedure either has not yet been automated or cannot be automated. `(The Clarin User Guide)`

3.3.2 Automatic vs. Manual

- Manual:
  - follow a linguistic theory
  - reliable (at least for the annotator!)
  - fastidious
- Automatic:
  - consistent
  - very fast
  - some rate of error

3.4 Brief overview of some software

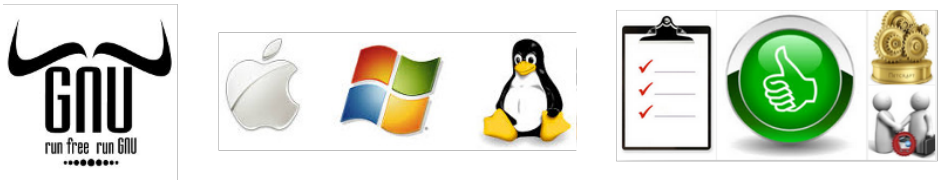
- In the following, we will briefly present some of the annotation software we already tested and validated to be part of the proposed methodology:
  - Praat

- Annotation Pro
- Elan
- SPPAS

3.4.1 Software requirements

- Most of them are:

  1. under the terms of the GNU Public License, and multi-platform,
  2. ease of use (GUI), with a tutorial and/or documentation,
  3. well-known in their communities, with publications and evaluations.



3.4.2 Praat: the analysis the software is designed for

- Praat is a tool for manually annotating sound files. It provides different visualizations of audio data – waveform or spectrogram display – and, among other, enables pitch contour and formant calculation and visualization.
- Annotations can be created on multiple layers, called tiers.

3.4.3 Praat: the type of data

- The annotation files are in several Praat-specific text formats.
- Interoperability: none!

3.4.4 Praat: screenshot

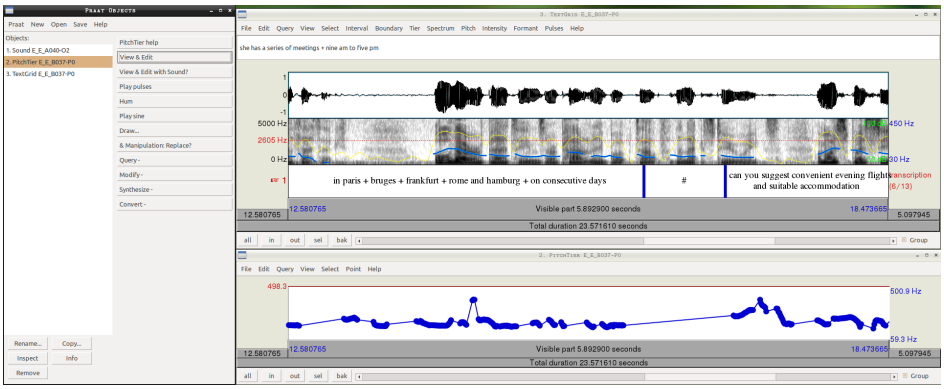


Figure 3.1: http://www.praat.org

3.4.5 Annotation Pro: the analysis the software is designed for

free download, Windows only

- Tool for annotation of audio and text files
- You can create many time-aligned annotation layers
- Workspace functionality enables comfortable file management, grouping files, opening previously stored file collections

- Graphical representation of the feature space is an innovative solution that enables using non-categorial features for the annotated recordings or texts
- Supports the design and conducting of Perception Tests

3.4.6 Annotation Pro: the type of data

- The annotation files are in a specific XML format
- Interoperability:
  - can import/export TextGrid, from Praat
  - can import trs, from Transcriber

3.4.7 Annotation Pro: screenshot

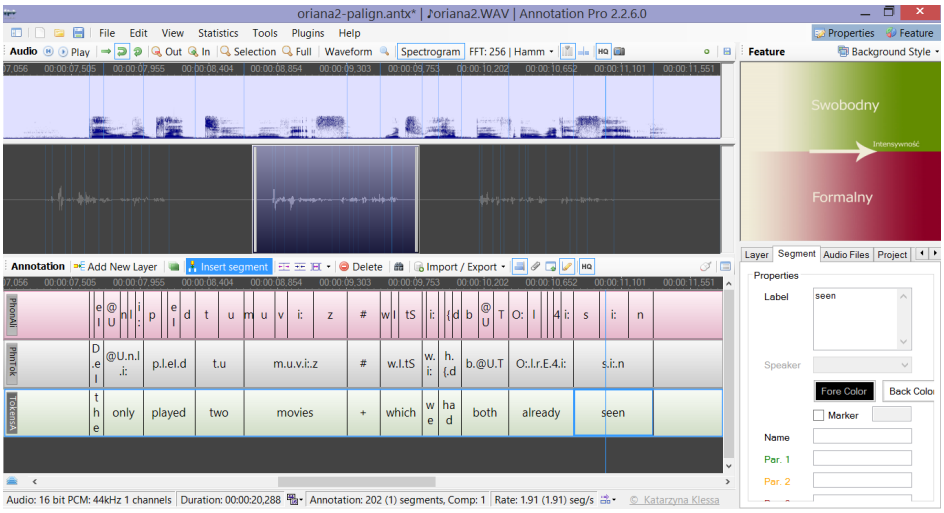


Figure 3.2: http://annotationpro.org/

3.4.8 Elan: the analysis the software is designed for

- Elan is a tool for the creation of complex annotations in video (and audio) resources.
- Annotations can be created on multiple layers, that can be hierarchically interconnected and can correspond to different levels of linguistic analysis.

3.4.9 Elan: the type of data

- The annotation files are in a specific XML format
- Interoperability: Annotation can be imported from and exported to a variety of other formats, including Praat-TextGrid.

3.4.10 Elan: screenshot

3.4.11 SPPAS: the analysis the software is designed for

- SPPAS is a free audio annotation tool that allows you to create, visualize and search annotations for audio data. It is able to produce **automatically speech segmentation annotations** from a recorded speech sound and its transcription. Some special features are also offered for managing corpora of annotated files.

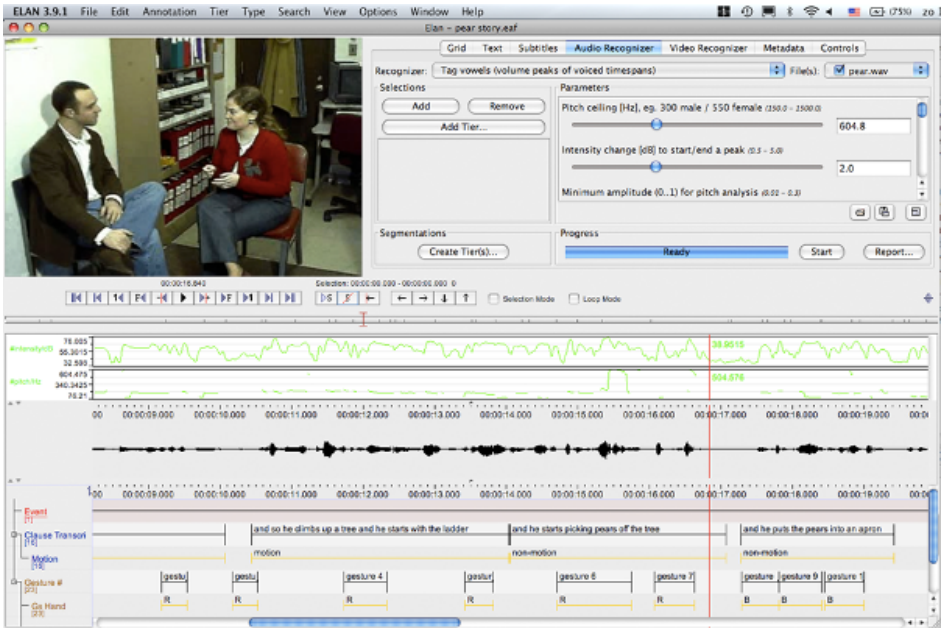


Figure 3.3: https://tla.mpi.nl/tools/tla-tools/elan/

3.4.12 SPPAS: the type of data

- The annotation files are in a specific XML format
- Interoperability: Annotation can be imported from and exported to a variety of other formats, including:
  - Praat: TextGrid, PitchTier, IntensityTier
  - Elan: eaf
  - Annotation Pro: antx
  - Phonedit: mrk
  - HTK: lab, mlf
  - Sclite: ctm, stm
  - subtitles: sub, srt
  - Transcriber: trs (import)
  - Anvil: anvil (import)
  - CSV

3.4.13 SPPAS: dedicated to automatic annotations

- Language-independent algorithms:
  - language-dependent resources
  - easy and fast to add a new language
- Fully-automatic or semi-automatic (with a procedure outcome report)
- Designed to be able to deal with spontaneous speech
- No limit of the corpus size

3.4.14 SPPAS: screenshot

3.5 Summary

- Introduction
- Selection of annotation software
- **Corpus development methodology**
- SPPAS
- Conclusion and references

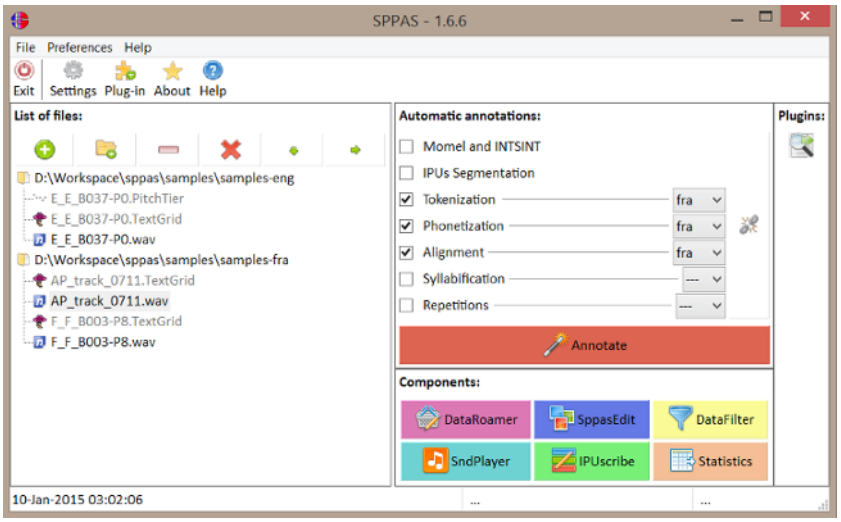


Figure 3.4: http://sldr.org/sldr000800/preview/

4

The annotation workflow

4.1 Information

- The proposed methodology is designed for the human analyst (mostly researchers in Linguistics).
- Therefore, we assume that the methodology is general enough to be useful for broad class of research applications.
- Different analytical domains - e.g. speech and gesture - and theoretical perspective require a rigorous organization of the annotation procedure.

4.2 Limitation

- The scope of the proposed workflow is broad and, therefore, complete coverage is challenging.
- It is very unrealistic to consider that human analyst can be removed from the process of annotation.

4.3 Which annotations (in general)?

> A very large number of dimensions have been annotated in the past on mono and multimodal corpora. To quote only a few, some frequent speech or language based annotations are speech transcript, segmentation into words, utterances, turns, or topical episodes, labeling of dialogue acts, and summaries; among video-based ones are gesture, posture, facial expression [...]. (Popescu-Belis, 2010)

Figure 4.1: Removing human from the process... a nod for whose who know!

4.3.1 Which annotations (in this tutorial)?

In this tutorial, we will report on:

1. IPUs segmentation (automatic)
2. Speech transcript (manual)
3. Phonemes and words segmentation (automatic)
4. Syllables segmentation (automatic)
5. Repetitions detection (automatic)
6. Morpho-syntax (automatic)
7. Momel and INTSINT (automatic)
8. Gestures (manual)

4.3.2 The annotation workflow: legend

4.3.3 The annotation workflow

4.3.4 The main principle is...

Garbage in, Garbage out.

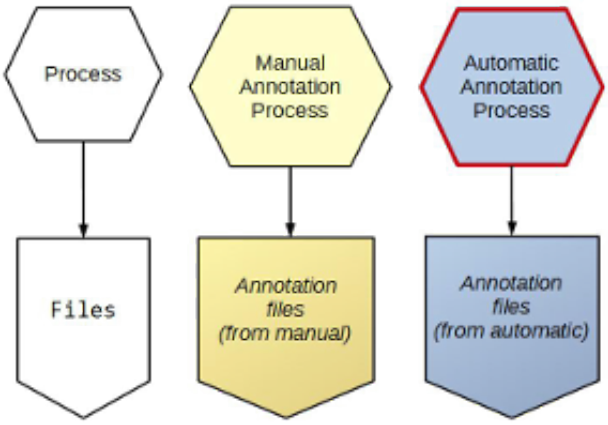


Figure 4.2: Legend of the annotation workflow

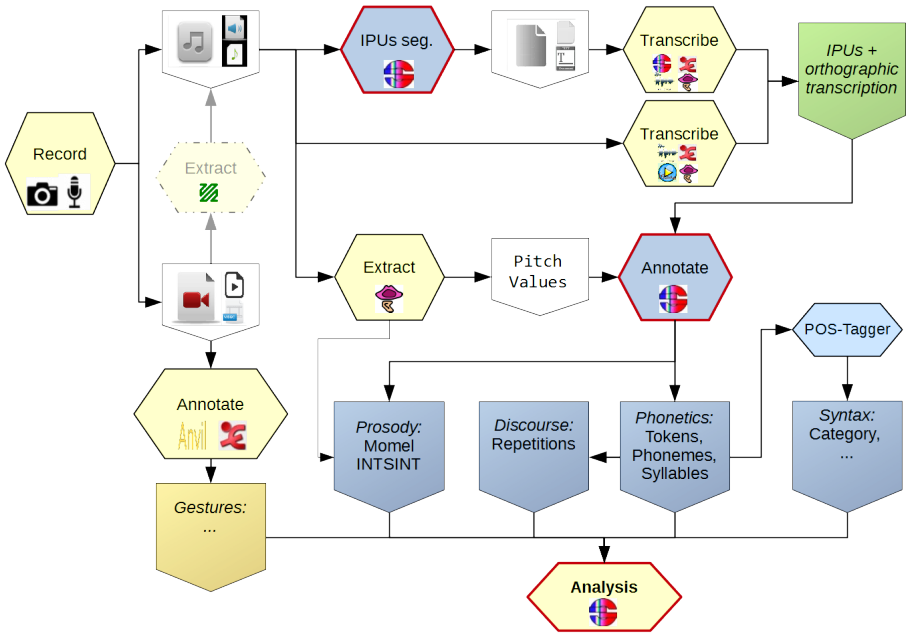
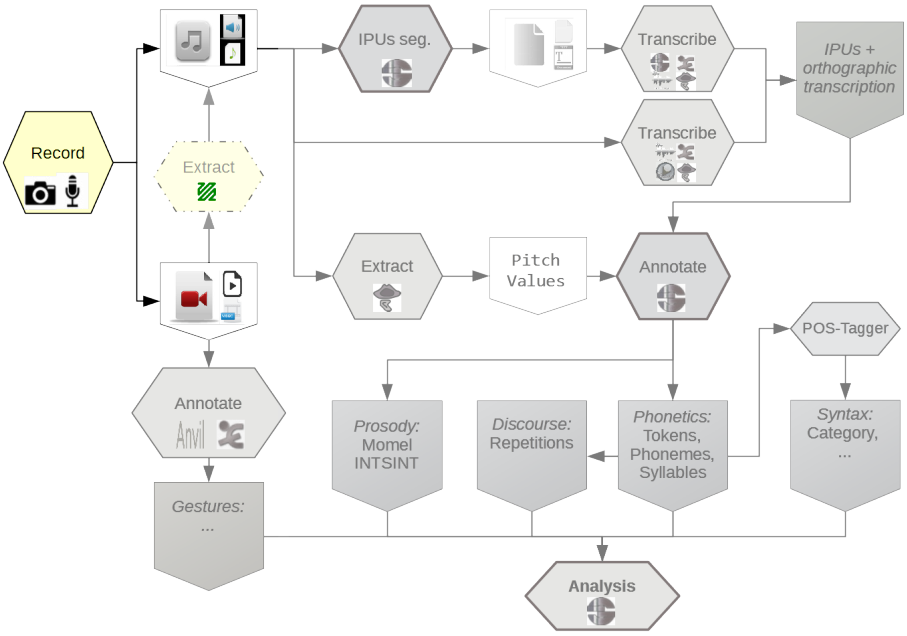


Figure 4.3: The annotation workflow



4.4 Record

4.4.1 Capturing and recording multimodal data

The capture of multimodal corpora requires complex settings such as instrumented lecture and meetings rooms, containing capture devices for each of the modalities that are intended to be recorded, but also, most challengingly, requiring hardware and software for digitizing and synchronizing the acquired signals. (Popescu-Belis, 2010)

4.4.2 Recording Audio and Video

- The resolution of the capture devices (microphones, framerate, file format, software) has a determining influence on the quality of the corpus, and so on the annotations.

- The number of devices is also important.
- Lack of standardization means that fewer researchers will be able to work with those signals.

4.4.3 Recording Audio: some advice

- One channel per speaker
- Anechoic room, or an environment with no/low noise
- Audio, for automatic annotation tools:
  - Any un-compressed file format, commonly WAV
  - 16000Hz is enough
- Audio, for manual annotation tools:
  - Any un-compressed file format
  - Most of the time 20000Hz is enough:
    - prefer 32000Hz/48000Hz if an high-quality is required

  Of course, provide 44100Hz

4.4.4 Recording Video: some advice

- Video file format:
  - refer to the annotation tool/software, and make tests!
  - provide compressed file formats
  - provide proprietary file formats
  - prefer H.264, it's a standard
  - prefer to record directly into the expected format (conversions are randomly good...)
- Take care of the lights (prefer LED)
- Pay attention to the noise the camera, the lights or the electricity power could generate

---

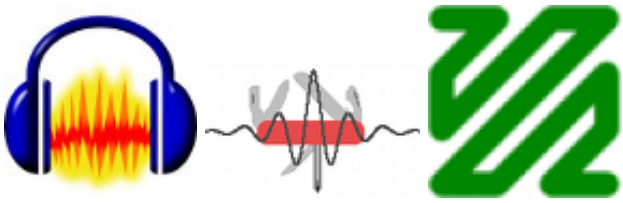
4.4.5 Synchronizing

- Synchronization of the signals is a crucial feature
- A regular "clap" (while recording) helps in this fastidious task (it's likely to "filming" the same clock on several signals).

4.4.6 Recommended tools/software

A short list of software we already tested and checked:

- audacity (audio) http://audacity.sourceforge.net/
- sox (audio) http://sox.sourceforge.net/
- ffmpeg (audio+video) https://www.ffmpeg.org/

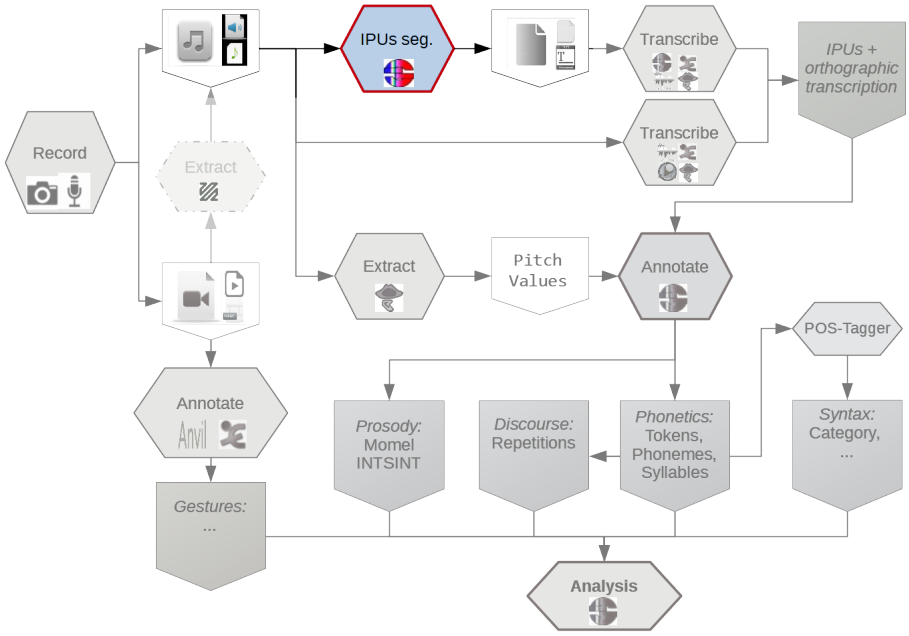


4.5 IPUs Segmentation

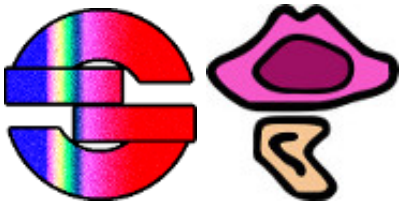
4.5.1 IPUs Segmentation: definition

- Automatic segmentation in Inter-Pausal Units
  - is also called Silence/Speech segmentation
- Parameters to define manually:
  - fix the minimum silence duration
  - fix the minimum speech duration
  - both values depend on:
    - the language
    - the speech style
- As results:
  - speech and silences are time-aligned and annotated automatically

---



4.5.2 IPUs Segmentation: software



- SPPAS is recommended
- A manual verification is recommended

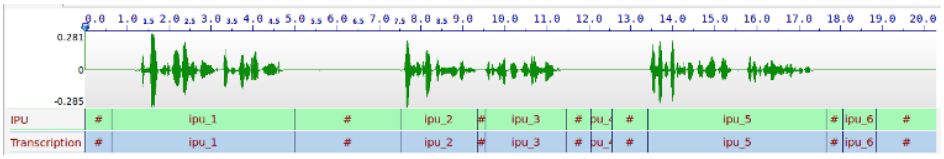
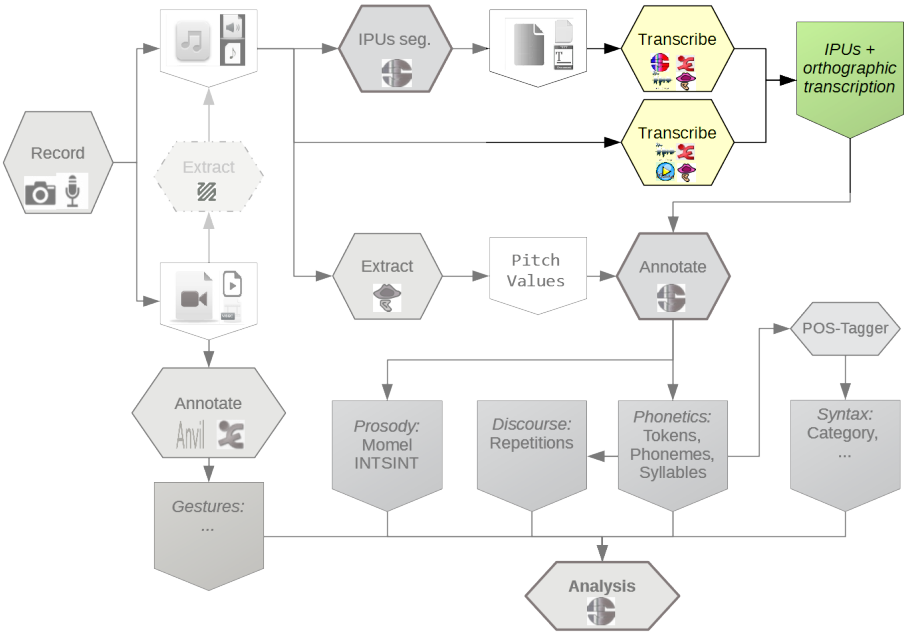


Figure 4.4: Example of IPUs segmentation: Silences are annotated with # and speech intervals are filled with ipu number

4.6 Orthographic Transcription



4.6.1 Orthographic Transcription

- An orthographic transcription is the minimum requirement for a speech corpus,
  - a better representation of pronunciation may be desired for most of research questions
- Orthographic transcription is at the top of the annotation procedure:
  - and remember: “Garbage in, Garbage out.”
- Orthographic transcription of spoken language presents considerable challenges.

4.6.2 Orthographic Transcription

- Speech may be annotated for:
  - phonemic transcription;
  - phonetic transcription taking into account details of pronunciation
    - allows a time-alignment at the phoneme-level
    - which is extended to time-alignment at word-level and syllable-level.
  - syntax analysis

4.6.3 Orthographic Transcription

- The better orthographic transcription implies:

  - the better phonetic transcription,
  - thus, the better time-alignment of phonemes,
  - thus, the better time-alignment of tokens,
  - thus, the better syllabification,
  - and so on...

- But, what is "the better" orthographic transcription?

  1. it's a representation of what is "perceived" in the signal
  2. it follows the convention the automatic system is requiring

---

4.6.4 Orthographic Transcription for spontaneous speech

- One of the characteristics of Spontaneous Speech is an important gap between a word's phonological form and its phonetic realizations.
- Specific realizations due to elision or reduction processes are frequent in spontaneous data.
- It also presents other types of phenomena such as:

  - non-standard elisions,
  - substitutions or addition of phonemes
  - noises, laughter, ...

- All of them intervene in the automatic system

4.6.5 Enriched Orthographic Transcription

- In speech (particularly in spontaneous speech), many phonetic variations occur:

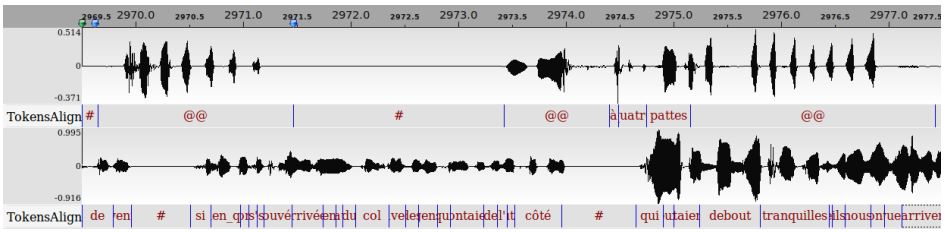
  - Some of these phonologically known variants are predictable

Transcription:	I	never	get	to	sleep	on	the	airplane
Phonetization:	ay	n.eh.v.e.r	g.eh.t g.ih.t	t.uw t.ix t.ax	s.l.iy.p	aa.n ao.n	dh.ax dh.ah dh.iy	eh.r.p.l.ey.n

  - but many others are still unpredictable (especially invented words, regional words or words borrowed from another language)

- The orthographic transcription must be enriched:

  - it must be a representation of what is "perceived" in the signal.



4.6.6 Enriched Orthographic Transcription

- In speech (particularly in spontaneous speech), many kind of events can occur like breathes, laughter, ...

4.6.7 Enriched Orthographic Transcription

- An EOT **must** include, at least:
  - Filled pauses
  - Short pauses
  - Repeats
  - Truncated words
  - Noises
  - Laughter
- An EOT must also include:
  - un-regular elisions
  - specific pronunciations
- An EOT may include:
  - all elisions

4.6.8 Enriched Orthographic Transcription: convention

- Any EOT must follow a convention
- The EOT is the input for automatic systems... and the transcription convention depends on the tool/software.
- So... you must read the documentation before starting to transcribe!

Figure 4.5: Train you first to transcribe and to use the annotation software!

4.6.9 SPPAS transcription convention

- truncated words, noted as a '-' at the end of the token string (an ex- example)
- noises, noted by a '*'
- laughs, noted by a '@'
- short pauses, noted by a '+'
- elisions, mentioned in parenthesis
- specific pronunciations, noted with brackets [example,eczap]
- comments are noted inside braces or brackets without using comma {this} or [this and this]
- liaisons, noted between '=' (an =n= example)
- morphological variants with <like,lie ok>
- proper name annotation, like $John S. Doe$

4.6.10 Transcription example 1 (Conversational speech)

- EOT:

  donc + i- i(l) prend la è- recette et tout bon i(l) vé- i(l) dit bon [okay, k]

- derived Standard orthograph:
  - donc il prend la recette et tout bon il dit bon okay
- derived Faked orthograph:
  - donc + i i prend la è recette et tout bon i vé i dit bon k

---

4.6.11 Transcription example 2 (Conversational speech)

- EOT:

  ah mais justement c'était pour vous vendre bla bla bla bl(a) le mec i(l) te l'a emboucané + en plus i(l) lu(i) a [acheté,acheuté] le truc et le mec il est parti j(e) dis putain le mec i(l) voulait

- Standard orthograph:
  - ah mais justement c'était pour vous vendre bla bla bla bla le mec il te l'a emboucané en plus il lui a acheté le truc et le mec il est parti je dis putain le mec il voulait
- Faked orthograph
  - ah mais justement c'était pour vous vendre bla bla bla bl le mec i te l'a emboucané + en plus i lu a acheuté le truc et le mec il est parti j dis putain le mec i voulait

---

4.6.12 Transcription example 3 (GrenelleII)

- EOT:

  euh les apiculteurs + et notamment b- on n(e) sait pas très bien + quelle est la cause de mortalité des abeilles m(ais) enfin il y a quand même + euh peut-êt(r)e des attaques systémiques

- Standard orthograph:
  - les apiculteurs et notamment on ne sait pas très bien quelle est la cause de mortalité des abeilles mais enfin il y a quand même peut-être des attaques systémiques
- Faked orthograph:
  - euh les apiculteurs + et notamment b on n sait pas très bien + quelle est la cause de mortalité des abeilles m enfin il y a quand même + euh peut-ête des attaques systémiques

---

4.6.13 Enriched Orthographic Transcription of 3 corpora

	CID	AixOx	Grenelle
Duration	143s	137s	134s
Number of speakers	12	4	1
Number of phonemes	1876	1744	1781
Number of tokens	1269	1059	550
Silent pauses	10	23	28
Filled pauses	21	0	5
Noises (breathes,...)	0	8	0
Laughter	4	0	0
Truncated words	6	2	1
Optional liaisons	4	2	5
Elisions (non stds)	60	21	34
Special Pron.	58	37	23

Figure 4.6: http://sldr.org/sldr000786

4.6.14 Orthographic Transcription... to sum up

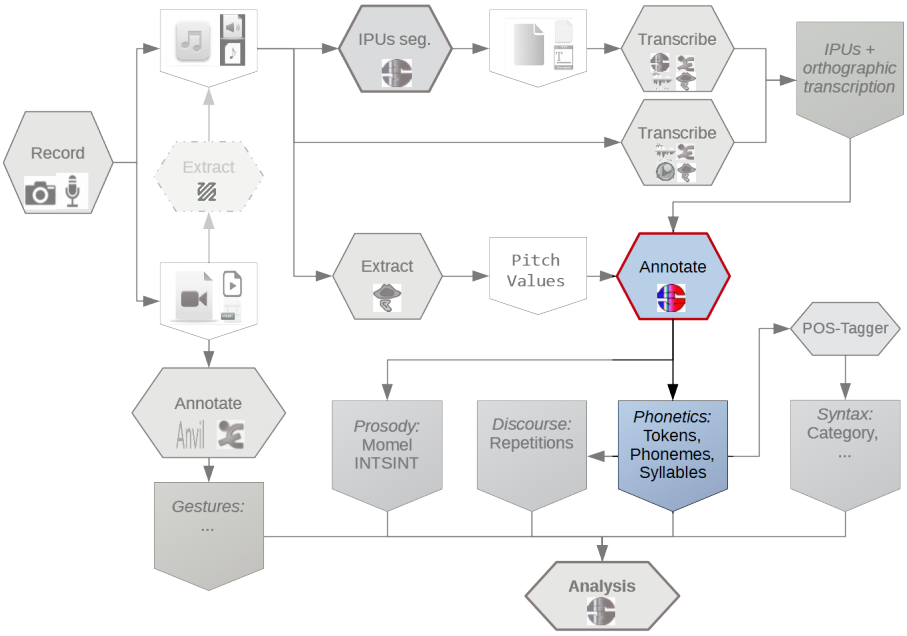
- An Enriched Orthographic Transcription is required
- The EOT of a corpus must follow a transcription convention
- Manual Standard orthographic transcription takes 15-20 minutes / minute of speech.
- Manual Enriched orthographic transcription takes 30-45 minutes / minute of speech.

The automatic systems must be adapted to deal with EOT

4.7 Phonemes/Tokens time-alignment

4.7.1 Phonemes and Tokens time-alignment

- A problem divided into 3 sub-tasks:
  1. tokenization
     - text normalization, word segmentation
  2. phonetization
     - grapheme to phoneme conversion
  3. alignment
     - speech segmentation



4.7.2 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

---

4.7.3 Tokenization in SPPAS

The main steps of the text normalization proposed in SPPAS are:

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

4.7.4 Tokenization in SPPAS

- From an EOT, SPPAS produces 2 outputs:
  - standard: the text normalization of the standard transcription,

  - faked: the test normalization of the faked transcription.

- Example:

  This is + hum... an enrich(ed) transcription {loud} number 1!

- standard: this is hum an enriched transcription number one
- faked: this is + hum an enrich transcription number one

`(Bigi 2011)`

4.7.5 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.

4.7.6 Phonetization in SPPAS

- SPPAS implements a dictionary based-solution
  - 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
- SPPAS implements a language-independent algorithm to phonetize unknown words.

`(Bigi 2013)`

4.7.7 Phonetization in SPPAS

By convention, spaces separate words, dots separate phones and pipes separate phonetic variants of a word. For example, the transcription utterance:

Input

the flight was twelve hours long and we really got bored

Output

dh.ax|dh.ah|dh.iy f.l.ay.t w.aa.z|w.ah.z|w.ax.z|w.ao.z t.w.eh.l.v
aw.er.z|aw.r.z l.ao.ng ae.n.d|ax.n.d w.iy r.ih.l.iy|r.iy.l.iy g.aa.t
b.ao.r.d

4.7.8 Impact of the Orthographic Transcription on automatic phonetization

- In (Bigi et al. 2012), we compared 3 types of OT:

  1. Standard orthographic transcription.
  2. Enriched 1: Std-OT + short pauses, various noises, laughter, filled pauses, truncated words, repeats.
  3. Enriched 2: Enriched 1 + elisions, particular pronunciations and unusual liaisons.

- Evaluations compare a reference phonetized manually to phonetizations obtained with SPPAS

4.7.9 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.

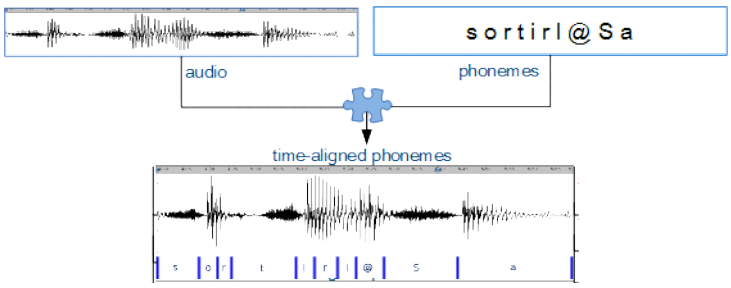


Figure 4.7: Time-alignment process

Manual alignment has been reported to take between 11 and 30 seconds per phoneme. (Leung and Zue, 1984)

4.7.10 How to perform Speech Segm. ?

1. 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
   - ...

4.7.11 How to perform Speech Segm. ?

2. Wrappers for such tool boxes:
   - Prosodylab-Aligner: python+HTK
   - P2FA: python+HTK
   - ...
3. Web-services:
   - WebMAUS
   - Train&Align
   - ...

4.7.12 How to perform Speech Segm. ?

- Packaged software

- user-friendly,
- with Graphical User Interface,
- with Command-line Interface,
- documentated,
- maintained,
- open-source,
- etc...

SPPAS (python+Julius), available for English, French, Italian, Spanish, Catalan, Polish, Japanese, Mandarin Chinese, Taiwanese, Cantonese

---

4.7.13 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):
  - tested on read speech
  - tested on conversational speech

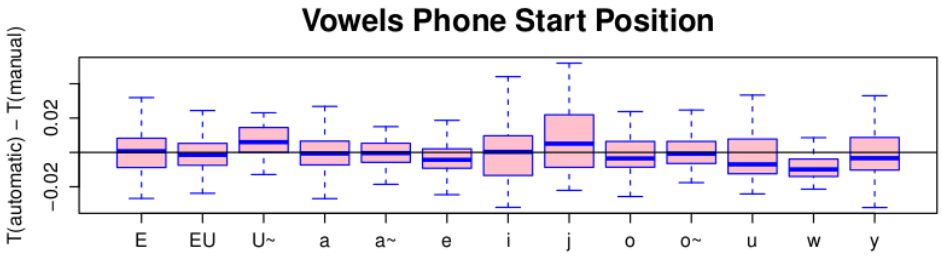
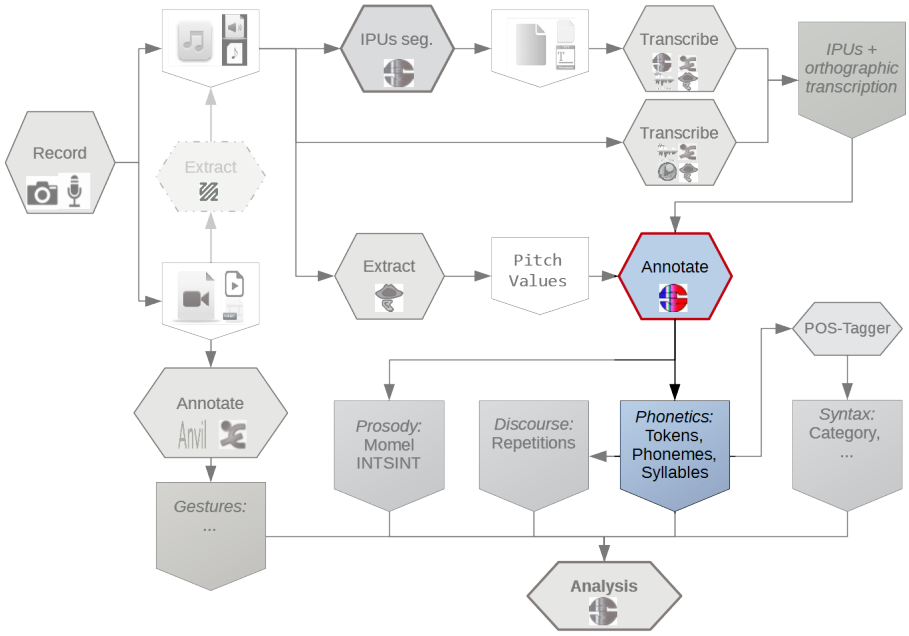


Figure 4.8: Results on vowels of French conversational speech

---

4.8 Syllables segmentation



4.8.1 Syllabification by SPPAS

- Automatic annotation
- A rule-based system
- Rules available for:
  - French
  - Italian
- 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.

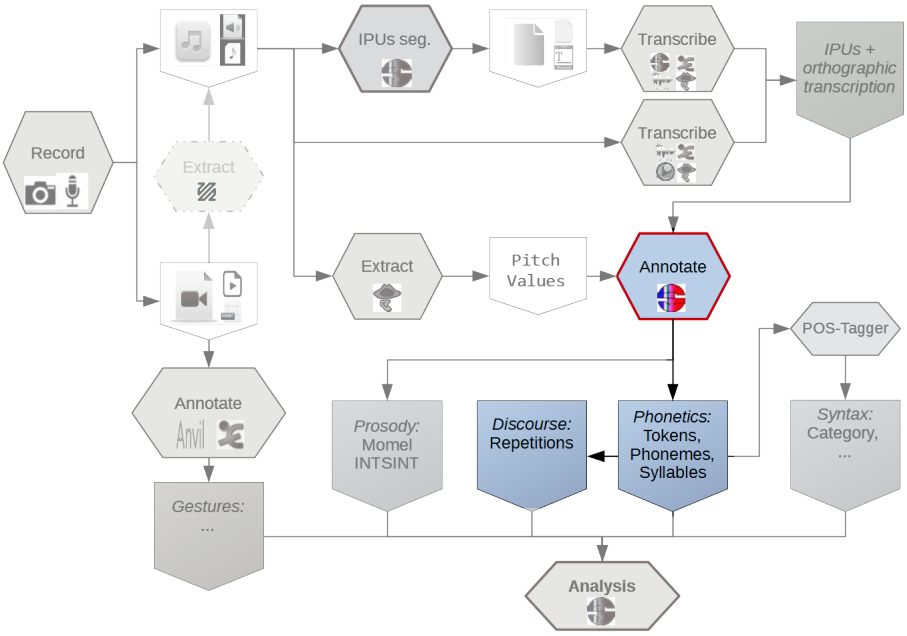
(Bigi et al. 2010)

4.8.2 Syllabification by SPPAS

- Phonemes are grouped into 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

Transcription	il expliquait pas vraiment ce qu'il y avait dedans
Phonemes	i l e k s p l i k e p a v r e m ã s k i j a v e d ã
Classes	V G V O F O L V O V O V F L V N V F O V V V F V O V
Syllables Auto	i . lek . spli . ke . pa . vre . mã . ski . ja . ve . dã
Syllables Expert1	i . lek . spli . ke . pa . vre . mã . ski . ja . ve . dã
Syllables Expert2	i . leks . pli . ke . pa . vre . mã . ski . ja . ve . dã

4.9 Repetitions detection



4.9.1 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

(Bigi et al. 2014)

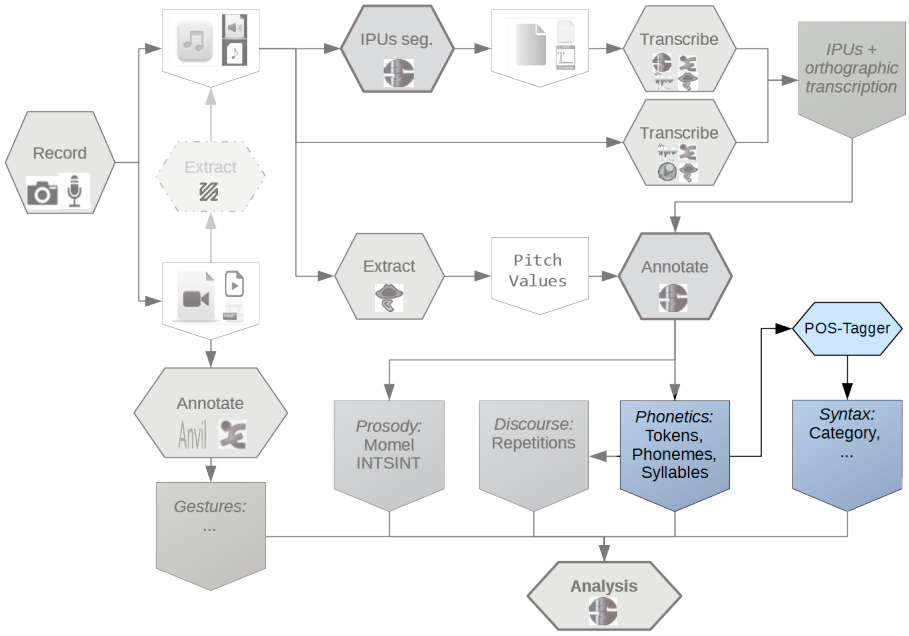
4.9.2 Repetitions

- Semi-automatic annotation performed by SPPAS
- SPPAS implements:

  - *self-repetitions*,
  - *other-repetitions* detection (CLI only).

- The system is based only on lexical criteria, from the time-aligned tokens (or lemmas)
- The system was used to propose a lexical characterization of OR: various statistics was estimated on the detected OR

4.10 Morpho-syntax



4.10.1 Morpho-syntax

- It is mostly dedicated to written text, in the NLP community
- A system must be adapted to deal with speech, particularly for conversational speech:
  - spoken data are time-aligned and we expect to get a time-aligned morpho-syntax!
  - the lexicon and the probabilities of tokens are different between written texts and speech, so they must be updated.
- At LPL, Stéphane Rauzy and G. De Montcheuil are proposing MarsaTag, for French:
  - http://sldr.org/sldr000841

4.10.2 Morpho-syntax: conversational speech vs map-task

- Annotated by MarsaTag, version 0.8

	CID	Aix MapTask
pronoun	**22,62%**	16,56%
verb	**17,62%**	15,25%
noun	11,37%	**15,34%**
adverb	10,67%	**12,39%**
determiner	8,70%	**11,75%**
preposition	7,80%	**11,75%**
conjunction	**7,77%**	4,71%
interjection	7,02%	7,14%
adjective	3,87%	3,67%
auxiliary	**2,56%**	0,77%

Figure 4.9: CID - conversational speech, versus Map-task speech

4.10.3 Example of Morpho-syntax in CID

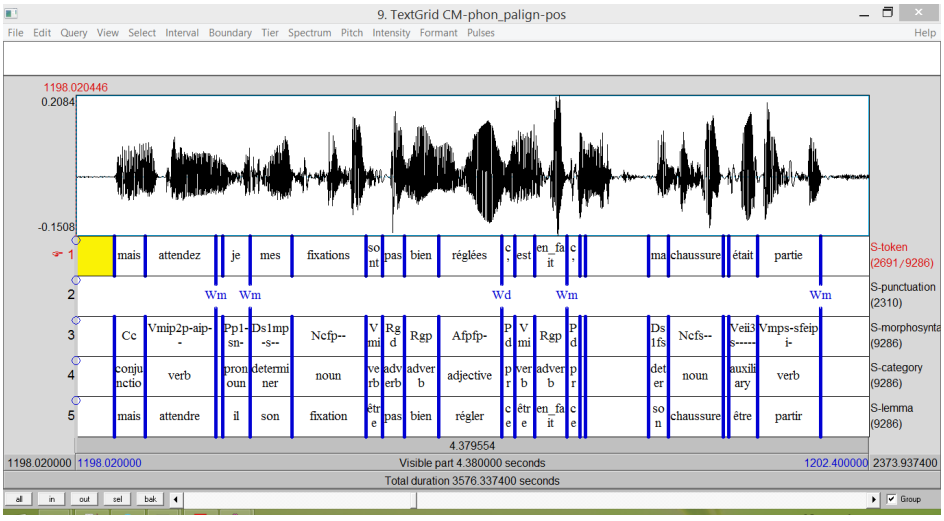
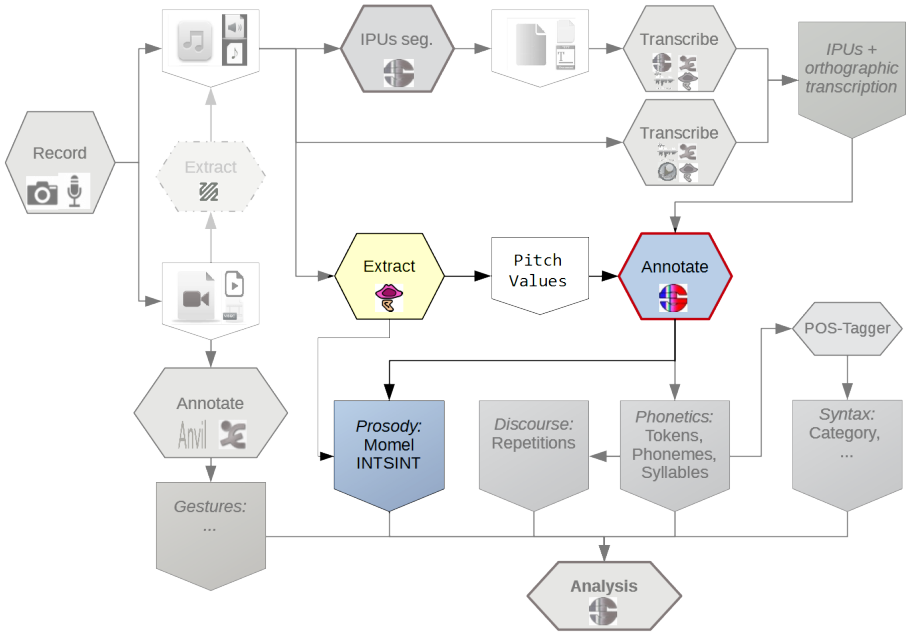


Figure 4.10: Example of time-aligned morpho-syntax on conversational speech

4.11 Momel and INTSINT



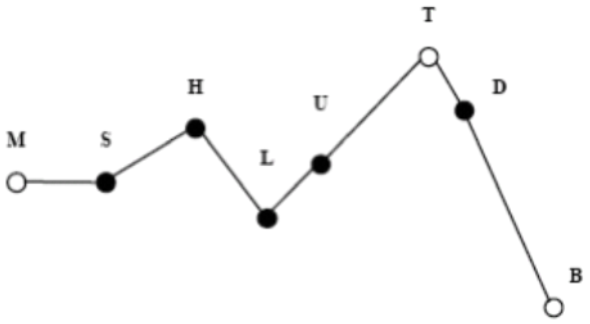
4.11.1 Momel and INTSINT

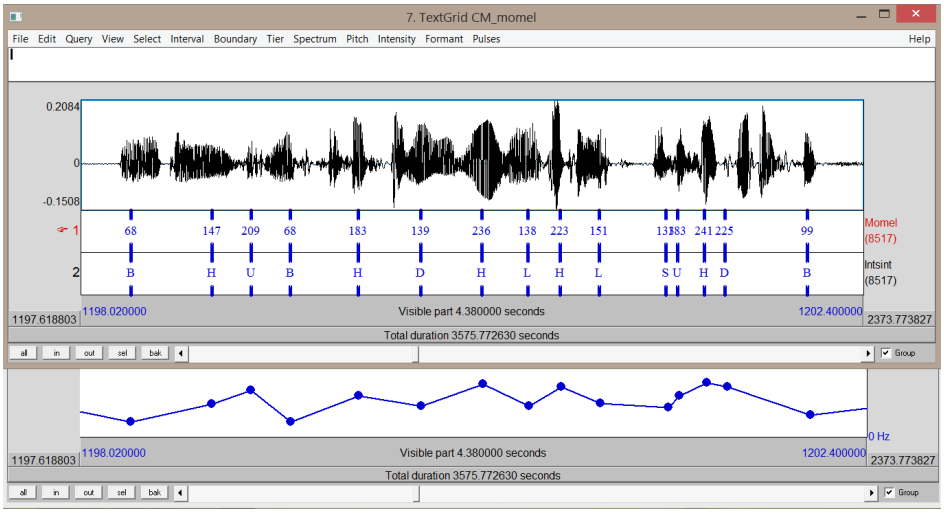
- Momel (modelling melody)
  - algorithm modelling raw fundamental frequency curves with a quadratic spline function
  - target F_0 Points
- INTSINT: an INternational Transcription System for INTonation
  - based on an inventory of minimal pitch contrasts found in published descriptions of intonation patterns
  - surface phonological structure
  - mapping from Momel target points to INTSINT tones

4.11.2 INTSINT

- Absolute tones: **T**(op) **M**(id) **B**(ottom)
- Relative tones: **H**(igher) **S**(ame) **L**(ower)
- Iterative relative tones: **U**(pstepped) **D**(ownstepped)

4.11.3 Example of Momel and INTSINT





4.11.4 Momel and INTSINT: software

- Momel and INTSINT are available:
  - as a Praat plugin, developped by Daniel Hirst
  - in SPPAS, developped by Brigitte Bigi

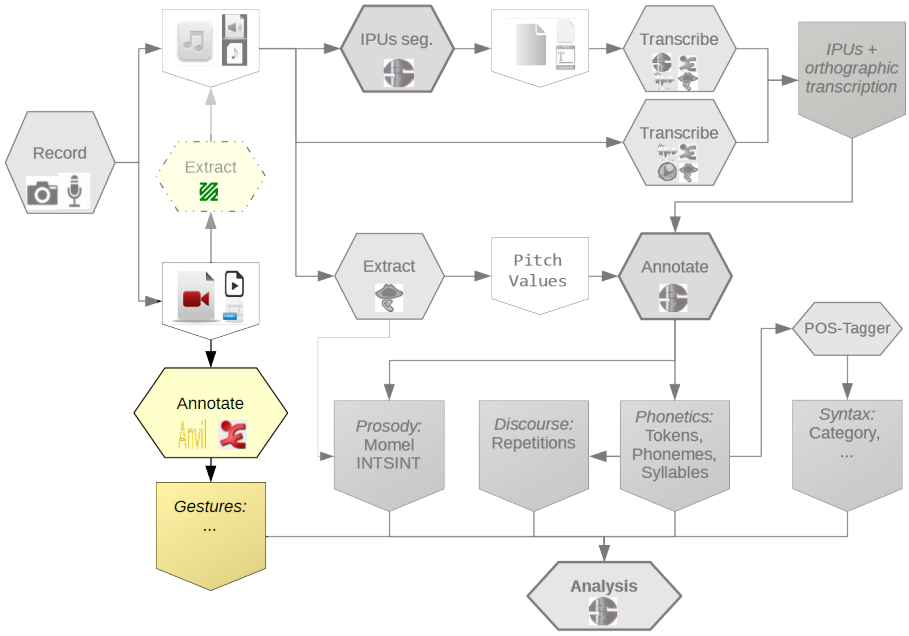
(Hirst and Espesser, 1993)

4.12 Gestures

4.12.1 Gestures: Annotation methodology

- http://discours.revues.org/8917

(Tellier 2014)



4.13 Summary

- Introduction
- Selection of annotation software
- Corpus development methodology
- **SPPAS**
- Conclusion

5

SPPAS

5.1 SPPAS: Automatic Annotation of Speech

- Brigitte Bigi
- http://sldr.org/sldr000800/preview/

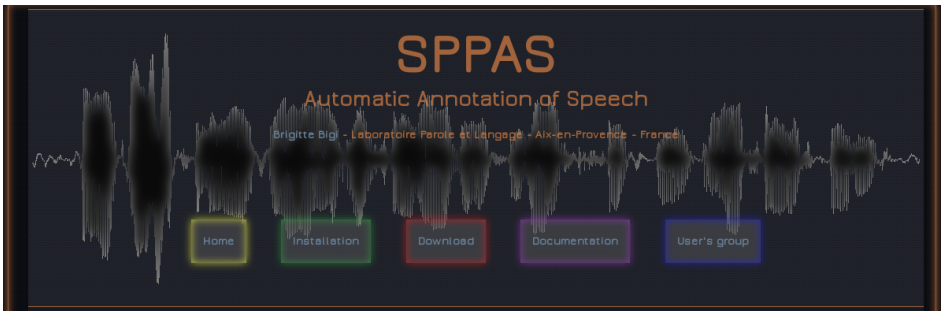


Figure 5.1: SPPAS web site

5.2 Install and update

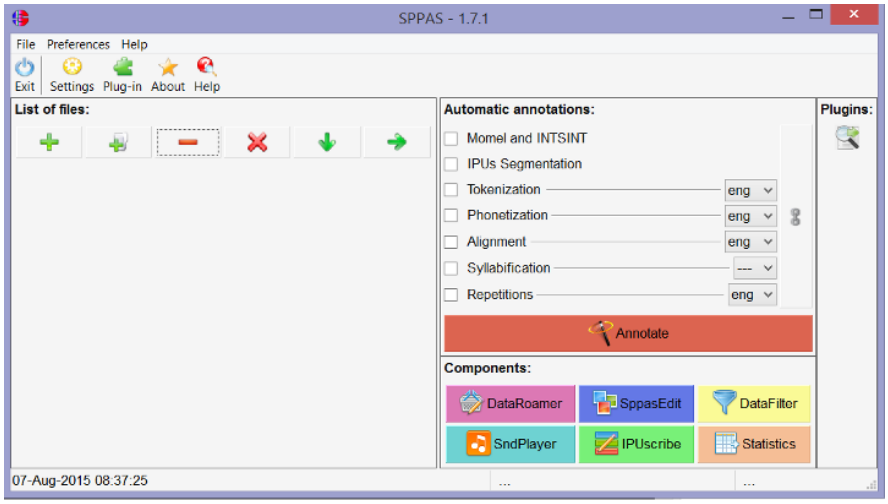
- Install:

  1. Follow *carefully* instructions of the installation page for dependencies:
     - python 2.7.x
     - wxpython > 3.0
     - julius CSR engine
  2. Download the last package (a zip file)
  3. Unzip on your computer

- Update SPPAS regularly:

  1. Put the old package into the Trash
  2. Download and unpack the new one

5.3 GUI usage

- Open the file explorer of your system
- Go to the SPPAS folder location
- Windows: doucle-click on the `sppas.bat` file
- MacOS / Linux: double-click on the `sppas.command` file
- 

5.3.1 GUI usage

- Click on the 'Add File' button
- Explore the `samples` folder and choose as many audio files as expected
- All files with the same name as the selected audio files will be added into the list
- Click (and/or ctrl+click) on some files in this list
- Choose what you want to do with your selection (a component, automatic annotations, plugin)

5.3.2 GUI usage

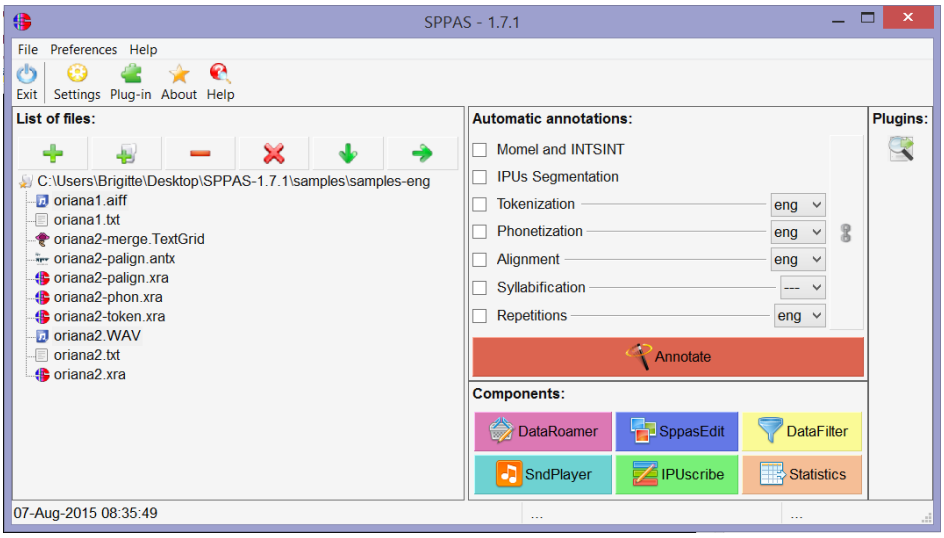


Figure 5.2: SPPAS main frame

5.4 Demo

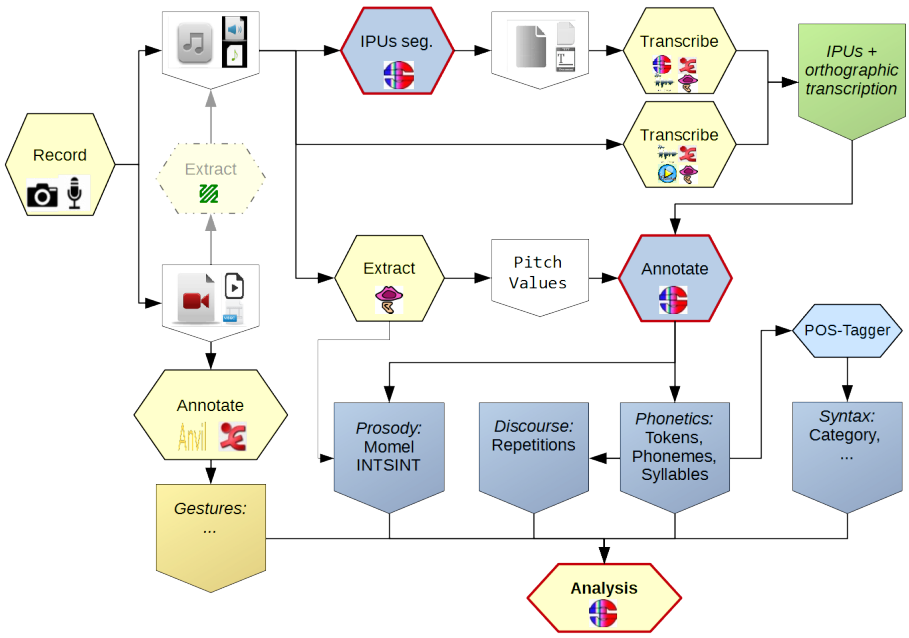
5.5 Summary

- Introduction
- Selection of annotation software
- Corpus development methodology
- SPPAS
- Conclusion

6

Conclusion

6.1 To sum-up



7

References

Abuczki, Ágnes, and Esfandiari Baiat Ghazaleh (2013). An overview of multimodal corpora, annotation tools and schemes. Argumentum, Hungria 1, no. 9: 86-98.

Bertrand, Roxane, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. Traitement automatique des langues 49, no. 3: 1-30.

Bigbee, Tony, Dan Loehr, and Lisa Harper (2001). Emerging requirements for multi-modal annotation and analysis tools. In INTERSPEECH, pp. 1533-1536.

Bigi, Brigitte, 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.

Bigi, Brigitte (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.

Bigi, Brigitte (2012). SPPAS: a tool for the phonetic segmentations of Speech. The eight international conference on Language Resources and Evaluation, Istanbul (Turkey), pages 1748-1755, ISBN 978-2-9517408-7-7.

Bigi, Brigitte , Pauline Péri, Roxane Bertrand (2012). Orthographic Transcription: Which Enrichment is required for Phonetization?, Language Resources and Evaluation Conference, Istanbul (Turkey), pages 1756-1763, ISBN 978-2-9517408-7-7.

Bigi, Brigitte and **Daniel Hirst** (2012). SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. Speech Prosody, Tongji University Press, ISBN 978-7-5608-4869-3, pages 19-22, Shanghai (China).

Bigi, Brigitte (2013). A phonetization approach for the forced-alignment task. 3rd Less-Resourced Languages workshop, 6th Language & Technology Conference, Poznan (Poland).

Bigi, Brigitte (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.

Bigi, Brigitte, Roxane Bertrand and Mathilde Guardiola (2014). Automatic detection of other-repetition occurrences: application to French conversational speech. 9th International conference on Language Resources and Evaluation (LREC), Reykjavik (Iceland), pages 2648-2652. ISBN: 978-2-9517408-8-4.

Bigi, Brigitte, Tatsuya Watanabe and Laurent Prévot (2014). Representing Multimodal Linguistics Annotated Data. 9th International conference on Language Resources and Evaluation (LREC), Reykjavik (Iceland), pages 3386-3392. ISBN: 978-2-9517408-8-4.

Chiarcos, Christian, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede (2008). A flexible framework for integrating annotations from different tools and tagsets. Traitement Automatique des Langues 49, no. 2: 271-293.

Gibbon, Dafydd, Inge Mertins and Roger Moore, eds. (2000). Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation. Dordrecht: Kluwer Academic Publishers.

Gibbon, Dafydd (2013). TGA: a web tool for Time Group Analysis, in D.J. Hirst & B. Bigi (Eds.) Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013. pp. 66-69.

Gibbon, Dafydd and Jue Yu (2013). How natural is L2 prosody? Proceedings of ICPhS 2015, Glasgow.

Gorisch, Jan, Corine Astésano, Ellen Gurman Bard, **Brigitte Bigi**, and Laurent Prévot (2014). Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. 9th International conference on Language Resources and Evaluation, ISBN 978-2-9517408-8-4, Reykjavik (Iceland)“, pages 2648-2652.

Hirst, Daniel (1987). La représentation linguistique des systèmes prosodiques : une approche cognitive. Thèse de Doctorat d'Etat (Habilitation Thesis), Université de Provence.

Hirst, Daniel and Robert Espesser (1993). Automatic Modelling Of Fundamental Frequency Using A Quadratic Spline Function. Travaux de l'Institut de Phonétique d'Aix, pages 75-85, vol. 85.

Hirst, Daniel and Di Cristo, Albert. (eds) (1998). Intonation Systems. A survey of Twenty. Languages. (Cambridge, Cambridge University Press). ISBN 0 521 39513 S (Hardback); 0 52139550 X (Paperback).

Hirst, Daniel, Albert Di Cristo and Robert Espesser (2000). Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (Ed.), Prosody: Theory and Experiment. Studies Presented to Gösta Bruce. (pp. 51–87). Kluwer Academic Pub.

Hirst, Daniel (2005). Form and function in the representation of speech prosody. Speech Communication, 46, 334–347.

Hirst, Daniel (2006). Review of John Coleman 2005. Journal of the International Phonetic Association. 198-200.

Hirst, Daniel (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modeling and coding intonation. In Proceedings of the XVIth International Conference of Phonetic Sciences, (paper 1443), pp 1233-1236.Saarbrücken, August 2007.

Hirst, Daniel (2015). ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis. In Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis (pp. 3-17). Springer Berlin Heidelberg.

Klessa, Katarzyna and **Dafydd Gibbon** (2014). Annotation Pro + TGA: automation of speech timing analysis. Proceedings of LREC 2014, Reykjavik. Paris: ELDA.

Leech, G. (1997). Introducing corpus annotation. In "Corpus Annotation: Linguistic Information from Computer Text Corpora", R. Garside, G. Leech & AM McEnery, ed.

Popescu-Belis, Andrei (2010). Managing Multimodal Data, Metadata and Annotations: Challenges and Solutions. Chapter 11, pages 187-199, IDIAP.

Rauzy, Stéphane, Grégoire De Montcheuil and Philippe Blache (2014). MarsaTag, a tagger for French written texts and speech transcriptions. Second Asia Pacific Corpus Linguistics Conference, Hong Kong.

Tellier, Marion (2014). Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés. Discours. Revue de linguistique, psycholinguistique et informatique, vol. 15.

Yu, Jue and **Dafydd Gibbon** (2012). Criteria for database and tool design for speech timing analysis with special reference to Mandarin, Oriental COCOSDA 2012, Macau (IEEEexplore Conf ID 21048).

Yu, Jue (2013). Timing analysis with the help of SPPAS and TGA tools, TRASP 2013, Aix-en-Provence.

Yu, Jue, **Dafydd Gibbon** and Katarzyna Klessa (2014). Computational annotation-mining of syllable durations in speech varieties. Proceedings of 7th Speech Prosody Conference, 20-23 May 2014. Dublin.