SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech

Brigitte Bigi

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

Abstract

The first step of most acoustic analyses unavoidably involves the alignment of recorded speech sounds with their phonetic annotation. This step is very labor-intensive and cost-ineffective since it has to be performed manually by experienced phoneticians during many hours of work.

This paper describes the main features of SPPAS, a software tool designed for the needs of automatically producing annotations of speech at the level of utterance, word, syllable and phoneme based on the recorded speech sound and its orthographic transcription. In other words, it can automatize the phonetic transcription task for speech materials, as well as the alignment task of transcription with speech recordings for further acoustic analyses.

Special attention will be given to the methodology implemented in SPPAS, based on algorithms which are as language-and-task-independent as possible. This procedure allows for the addition of new languages quickly and for the adaptation of this tool to the user's specific needs. Consequently, the quality of the automatic annotations is largely influenced by external resources, and the users can modify the process as needed. In that sense, phoneticians need automatic tools and these tools can be significantly improved by phonetician input.

Keywords: automatic, annotation, speech segmentation, multilingual, methodology

1 Introduction

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). Annotation of speech recordings is relevant for many sub-fields of linguistics such as phonetics, prosody, gesture analysis or discourse studies. Corpora are annotated with detailed information at various linguistic levels, often with the use of specialized annotation software. As *large multimodal* corpora become prevalent, new annotation and analysis requirements are emerging. In order to be useful for purposes such as qualitative or quantitative analyses, the annotations must be time-synchronized (time-aligned). Temporal information makes it possible to describe behaviour or actions of different subjects that happen at the same time, and time-analysis of multi-level annotations can reveal levels of linguistic structures. Generally, “different annotation tools are designed and used to annotate the audio and video contents of a corpus that can later be merged in query systems or databases” (Abuczki and Baiat Ghazaleh, 2013). A number of software programs for manual annotation and analysis of audio and/or video recordings are available such as Transcriber (Barras et al. 2001), Praat (Boersma

and Weenink 2001), or Elan (Wittenburg et al. 2006), to name but just some popular ones that are both open-source and multi-platform.

In the past, phonetic study was mostly based on limited data. Currently, phonetic models are often expected to be built based on the acoustic analysis of large quantities of speech data supported with valid statistical analyses. The first step of most acoustic analyses unavoidably involves the alignment of recorded speech sounds with its phonetic annotation. This step is very labor-intensive and cost-ineffective since it has to be performed manually by experienced phoneticians requiring many hours of work. For speech engineers, this labor-intensive task can be assisted by computer programs. A number of free toolkits are currently available which can be used to automate the task, including the HTK Toolkit (Young and Young 1993), Sphinx (Lamere et al. 2003), or Julius (Lee et al. 2001). In recent years, the SPPAS software tool has been developed to automatically produce "annotations which include utterance, word, syllabic and phonemic segmentation from a recorded speech sound and its transcription" (Bigi 2012). In other words, this software can automatize the phonetic transcription task for speech materials, as well as the alignment task of matching transcriptions to the speech recordings for further acoustic analyses. SPPAS includes resources for various languages such as English, French, Italian, Spanish, and Mandarin Chinese. An important feature is that SPPAS is specifically designed to be used directly by linguists (not necessarily skilled in programming) in conjunction with other tools for the analysis of speech. It is a *free software*, as defined by Richard Stallman (2002), and distributed under the terms of the GNU Public License.

Modern technology gives linguists the means of refuting theories and models with large quantities of language data. In order to efficiently use annotation software, particularly for automatic annotations, a rigorous methodology is necessary. Section 2 of this paper presents how to collect a large set of time-aligned annotations for various domains or levels: orthographic transcription (time-aligned at the level of inter-pausal units), phonetics (words, syllables, phonemes), prosody (Momel and INTSINT), morpho-syntax (categories, groups), discourse (repetitions) and gestures. Some are annotated manually and most of them are generated automatically. The main features of SPPAS are presented together with the basic guidelines for its integration within such a framework. Section 3 describes the automatic annotations implemented in SPPAS, with algorithms as language-and-task-independent as possible. This allows adding new languages with a significant reduction of time compared to the development of such tools from scratch, because adding a new language in SPPAS only consists of adding the resources related to the annotation (like lexicons, dictionaries, models, sets of rules, etc). Consequently, the quality of the automatic annotations is largely influenced by such resources, and phoneticians can contribute to improve them.

2 Introducing SPPAS in a corpus construction methodology

This section illustrates the kind of process for development of a corpus that contains rich and broad-coverage of multimodal/multi-level annotations. This involves a rigorous framework to ensure compatibilities between accurate annotations and time-saving methodologies. Indeed, "when multiple annotations are integrated into a single data set, inter-relationships between the annotations can be explored both qualitatively (by using database queries that combine levels) and quantitatively (by running statistical analyses or machine learning algorithms)" (Chiarcos 2008). The expected result is time-aligned data, for all annotated levels including phonetics, prosody, gestures, syntax, discourse (cf. Figure 1). The

wide range of annotations is costly to collect and annotate, both in terms of time and money. Consequently, each annotation that *can* be done automatically *must* be done automatically, because revising is expected to be less time-consuming and easier than annotating, as shown for example by the use of SPPAS in Yu (2013). Fortunately, the current state-of-the-art in computational linguistics allows many annotation tasks to be semi- or fully- automated. Unfortunately, the lack of interoperability between automatic annotation tools/data and manual annotation tools/data is still a challenge. Thus, despite the advances that have been achieved for annotating and analyzing corpora, many annotation frameworks and/or models for the construction and analysis of multimodal data continue to rely on "low-tech" and/or manual technologies.

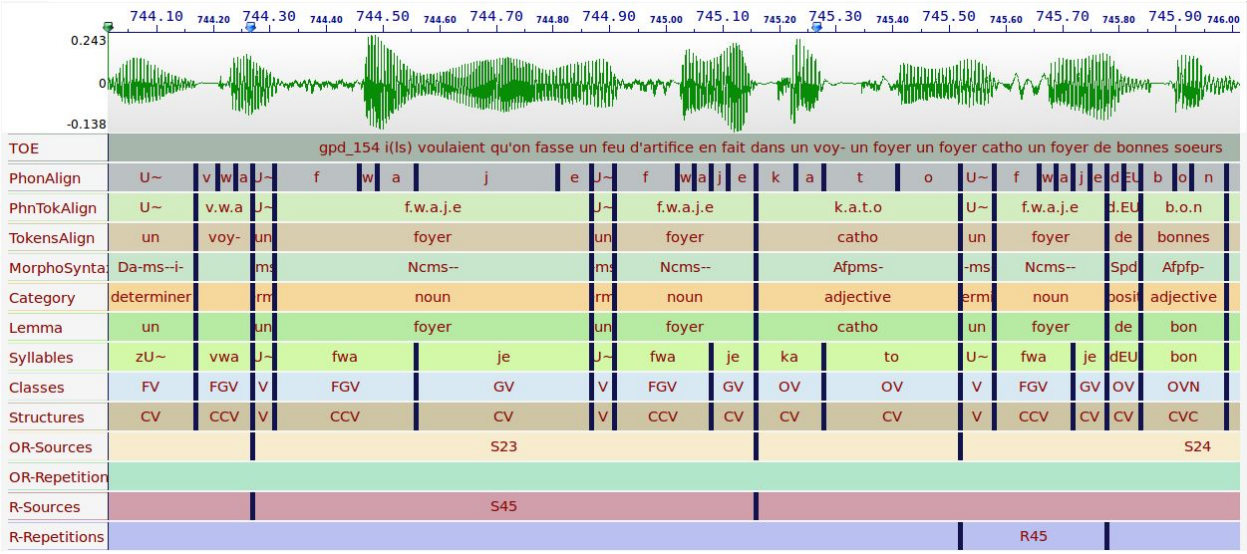

Figure 1: A selection of multi-level annotations based on the speech signal. The tier "TOE" is the enriched orthographic transcription and it was manually annotated. The other tiers were automatically annotated by SPPAS and MarsaTag (Rauzy 2014) software.

In recent years, many annotation software/tools have become available for annotation of audio-video data. For a researcher looking for an annotation software tool, it might be difficult to select the most appropriate one. The choice of the software determines the annotation framework and that will be utilized and this process should be done carefully and *before* the creation of the corpus. To decide about usefulness and usability of a software, it is advisable to consider the issues listed below.

- The software license: the preference is for free and open source software. Even if a user can personally afford to pay for a license, he/she may wish to share his/her methodology with other students or researchers who cannot afford to buy it.
- The ease of use: the first, preference is for multi-platform software. Different scientific communities tend to use MacOS, Windows or Unix platforms. Multi-platform software makes sharing between such communities much easier. Secondly, usable software is preferred. A need to request help from an engineer each time a user needs to use a piece of software may pose a serious limitation.
- The strengths/weaknesses for specific annotation purposes. Users should investigate if the software has been found to be reliable and is likely to improve the efficiency of annotation

workflow, by either accelerating the work or enabling one to deal with more extensive data, or both.

- The type of data or analysis the tool/software is specifically designed to complete.
- The software compatibility with other annotated data, i.e. the availability of files to be imported/exported from/to several other data formats.
- 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.

In the following part of this section, we very briefly introduce selected annotation software programs that were included as part of the proposed annotation methodology: Praat, Elan and SPPAS.

Praat is a tool for manually annotating sound files. It provides different visualizations of audio data - waveform or spectrogram display - and, among other things, enables pitch contour as well as formant calculation and visualization. The annotation files are in several Praat-specific ASCII formats, but Praat doesn't support any import or export to other formats. Fortunately, Praat-TextGrid file format is well-known in the community and external converters exist.

Elan is a tool for the creation of complex annotations for 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. It also includes an advanced search system. The annotation files are in a specific XML format, and Elan can import from and export to a variety of other formats, including Praat-TextGrid.

SPPAS is an annotation software that allows one to *automatically* create, visualize and search annotations of audio data. In fact, the analysis of the phonetic entities of speech nearly always requires the alignment of the speech recording with a phonetic transcription of the speech. This task is extremely labor-intensive - it may require several hours even for an experienced phonetician to transcribe and align manually a single minute of speech. It is thus obvious that transcribing and aligning several hours of speech by hand is not generally something which can be accomplished with ease. Therefore, among others, SPPAS includes automatic segmentation of speech. It offers a fully-automatic or semi-automatic annotation process, with a procedure outcome report to help the user in understanding particular steps. Some special features are offered in SPPAS for managing corpora of annotated files; e.g., a component to filter multi-level annotations (Bigi and Saubesty, 2015). Some other components are dedicated to the analysis of time-aligned data; as for example to estimate descriptive statistics, a version of Time Group Analyzer (Gibbon 2013), etc. SPPAS annotation files are in a specific XML format, and annotations can be imported from and exported to a variety of other formats, including Praat (TextGrid, PitchTier, IntensityTier), Elan (eaf), Transcriber (trs), Annotation Pro (antx) (Klessa et al. 2013, Klessa 2015), Phonedit (mrk) (Teston et al. 1999), Sclite (ctm, stm), HTK (lab, mlf), subtitles formats (srt, sub) and CSV files. SPPAS can be used either with a Command-line User Interface or a Graphical User Interface as shown in Figure 2. So, there's no specific difficulty when using this software. The only potential brake on its usage is the need to integrate it in a rigorous methodology for the corpus construction and annotations.

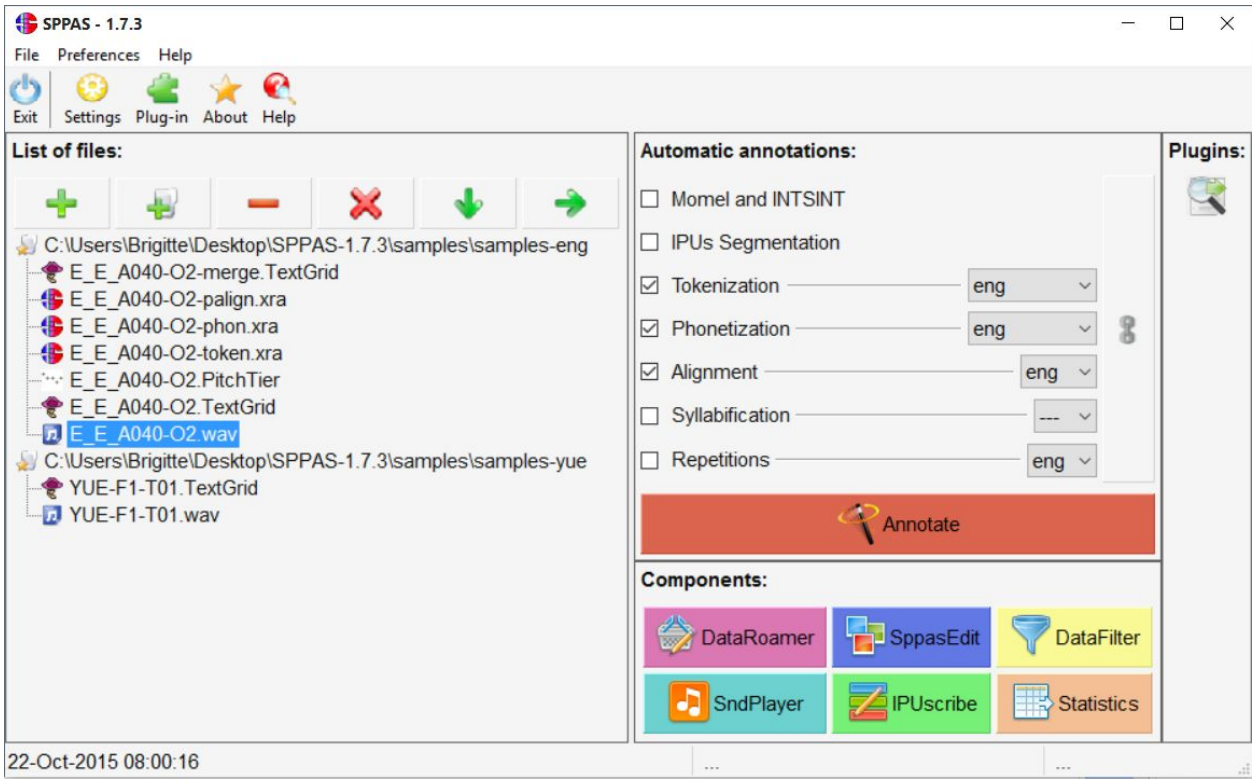

Figure 2: SPPAS Graphical User Interface. The left part indicates the list of files to work with; the middle part displays the functionalities of SPPAS (top: the whole list of automatic annotations; bottom: a set of 6 components provided to manage annotated data) and right part is dedicated to plug-ins (only one on this picture).

The kind of process for obtaining rich and broad-coverage of multimodal/multi-levels annotations of a corpus is illustrated in Figure 3. It describes each step of corpus creation and annotation workflow. This Figure must be read from top to bottom and from left to right, starting with the recordings and ending with the analysis of annotated files.

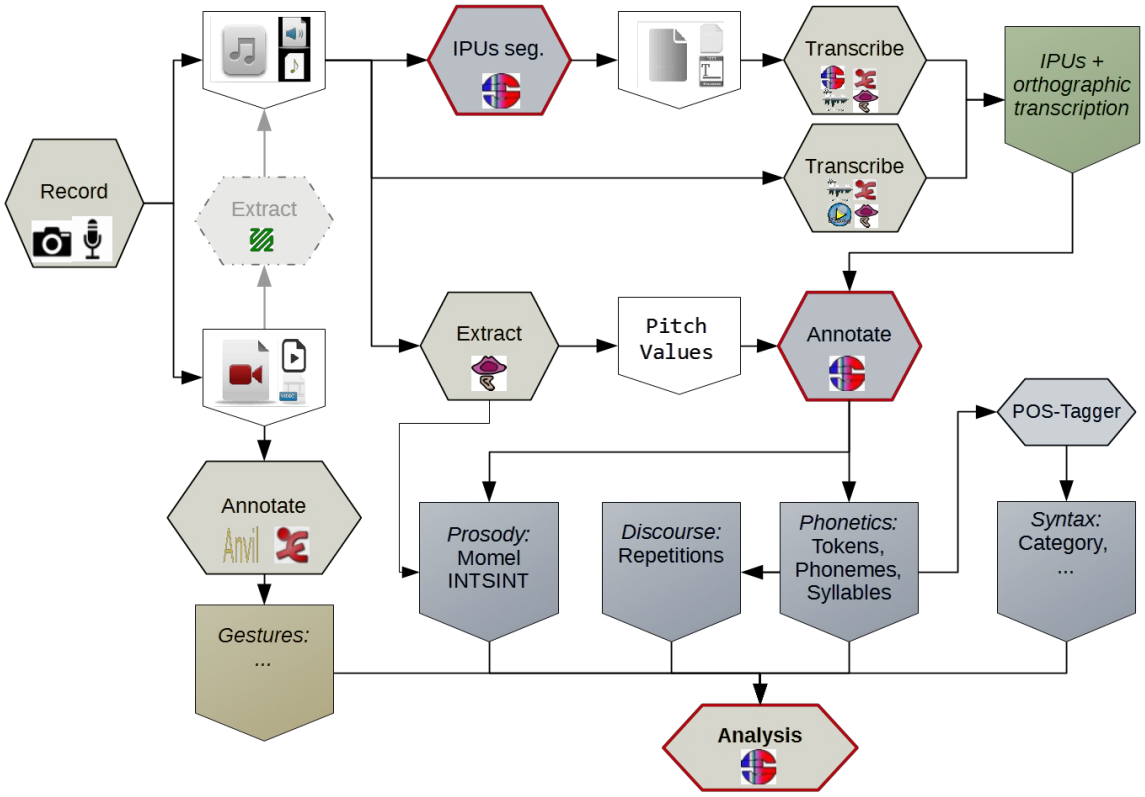

Figure 3: A multi-level corpus creation and annotation workflow. Yellow boxes represent manual annotations, blue boxes represent automatic ones.

After recording speech samples, the first step to perform is **IPUs segmentation**. IPUs (Inter-Pausal Units) are blocks of speech bounded by silent pauses of more than *X* ms (the *X* duration depends on the language; for French, the duration of 200 ms is commonly used), and time-aligned on the speech signal. IPUs segmentation should be verified manually. The outcome of this automatic procedure depends on the quality of the recording: the better the quality, the better IPUs segmentation.

Orthographic transcription is often the minimum obligatory requirement for a speech corpus, as it is the entry point for most of the automatic annotations, including automatic speech segmentation. As a consequence, high quality orthographic transcription implies:

- high quality phonetic transcription,
- thus, high quality time-alignment of phonemes and tokens,
- thus, high quality syllabification,
- and so on.

The question then arises: what is "the better" orthographic transcription method? First, one of the characteristics of speech is the important gap between a word's phonological form and its phonetic realizations. Specific realizations due to elision or reduction processes often occur and the same happens for other types of phenomena such as non-standard elisions, substitutions or addition of phonemes, noises, and laughter. Numerous studies have been carried out on prepared speech, such as broadcast news.

However, conversational speech refers to a more informal activity, in which participants constantly need to manage and negotiate turn-taking, topic, etc. "on line" without any preparation which results in an even greater number and wider variety of non-standard events. Table 1 reports on the amount of such phenomena taken from three manually annotated samples of the following French corpora:

1. AixOx, read speech of short texts (Herment et al. 2013);
2. Grenelle II, a discourse at the French National Assembly (Bigi et al. 2012);
3. CID - Corpus of Conversational Data, spontaneous dialogs (Bertrand et al. 2008).

	AixOx	**Grenelle II**	**CID**
Duration of the samples	137s	134s	143s
Number of speakers	4	1	12
Number of phonemes	1744	1781	**1876**
Number of tokens	1059	550	**1269**
Short silent pauses	23	**28**	10
Filled pauses	0	5	**21**
Noises (breathes, …)	**8**	0	0
Laughter	0	0	**4**
Truncated words	2	1	**6**
Optional liaisons	2	**5**	4
Elisions (non standard)	21	34	**60**
Specific pronunciations	37	23	**58**

Table 1: Description of events in three different corpora available at http://sldr.org/sldr000786

These events may create obstacles for the automatic annotation process. Thus, SPPAS includes the support of an Enriched Orthographic Transcription (EOT). Hese, transcribers are asked to indicate: filled pauses, short pauses, repeats, truncated words, noises, laughter, irregular elisions and specific pronunciations. These specific phenomena have a direct influence on the automatic phonetization procedure as shown in Bigi (2012).

The **Phonetics (Tokens, Phonemes, Syllables)** component of the workflow involves the process of taking the phonetic transcription text of an audio speech segment, like IPUs, and determining where particular phonemes occur in this speech segment. In SPPAS, this problem is clearly divided into three sub-tasks: Task 1 is tokenization, also called text normalization, Task 2 is phonetization, also called grapheme to phoneme conversion, and Task 3 is time-alignment, which is the speech segmentation task itself. All three sub-tasks are fully-automatic, but each annotation output can be manually checked if desired (a semi-automatic mode). The current version of SPPAS (1.7.4) includes data and models for: French, English, Italian, Spanish, Catalan, Portuguese, Polish, Mandarin Chinese, Cantonese, Taiwanese and Japanese. The time-alignment of tokens (usually words) can be automatically derived from the time-alignment of phonemes. Afterwards, the time-alignment of *syllables* is derived from the time-alignment of phonemes using a rule-based system (Bigi et al. 2010).

In the **Discourse** domain, as shown in Figure 3, the time-alignment of tokens can also be used by SPPAS to automatically identify self-repetitions and other-repetitions (OR). This system is based only on

lexical criteria to determine whether a token (only word in that case) is repeated or not. A set of rules are then fixed to filter such occurrences and to select only the relevant ones (Bigi et al. 2014). This system was used to propose a lexical characterization of OR: various statistics were estimated on the detected OR from CID corpus. It was also used to analyze if the same speech implies the same or different gestures in Tellier et al. (2012).

In the **Syntax** domain, a stochastic parser can be adapted to automatically generate morpho-syntactic and syntactic annotations. Actually, it must be adapted in order to account for the specifics of speech analysis, and to take time-aligned tokens as input. For French, MarsaTag (Rauzy 2014) is available and can be used as a plugin of SPPAS.

The **Prosody** domain can also be investigated and included as part of the framework. Momel (Hirst and Espesser 1993) is an example of a freely available algorithm for automatic modeling of fundamental frequency (F0) curves using a technique called asymmetric modal quadratic regression. This technique makes it possible to factor an F0 curve into two components by an appropriate choice of parameters:

1. a macroprosodic component represented by a quadratic spline function defined by a sequence of target points <ms,Hz>.
2. a microprosodic component represented by the ratio of each point on the F0 curve to its corresponding point on the quadratic spline function.

INTSINT (an INternational Transcription System for INTonation) assumes that pitch patterns can be adequately described using a limited set of tonal symbols, T, M, B, H, S, L, U, D (standing for: Top, Mid, Bottom, Higher, Same, Lower, Up-stepped, Down-stepped respectively). Each one of these symbols characterizes a point on the fundamental frequency curve. Momel and INTSINT are tools enabling automatic annotations and are available as a Praat plug-in (Hirst 2007), and re-implemented within SPPAS.

Gestures annotation can also play an important role in an annotation workflow, by reflecting the multimodal aspects of speech communication, however, this factor will not be described further in this paper. One can refer to Tellier (2014) for methodological insight into gesture annotation.

To sum up, this section presented a methodology for the annotation of recordings, based on both manual annotations and on annotations produced automatically with SPPAS, as illustrated in Figure 3. This methodology was established in the annotation of the CID - Corpus of Interactional Data (Bertrand et al. 2008, Blache et al. 2010), and SPPAS was initially created to generate annotations only on the level of Phonetics. Subsequently, several other corpora were created using SPPAS in the context of various projects, e.g.: Amennpro (Herment et al. 2013), Cofee (Gorish 2014), Multiphonia (Alazar et al. 2012), Typaloc (Bigi et al. 2015), and Variamu (Bigi and Fung 2015). In order to meet new expectations and new project requirements, SPPAS was improved and extended with new functionalities and components. The proposed methodology has demonstrated flexibility as well as effectiveness and reliability in the demanding, real-world situations of corpora creation.

3 SPPAS: multi-lingual approaches

3.1 Text normalization

The first task faced by any Natural Language Processing system is the conversion of input text into a linguistic representation. Digital written texts contain a variety of "non-standard" entry types such as digit

sequences, acronyms and letter sequences in all capitals, mixed case words, abbreviations, Roman numerals, URL's and e-mail addresses. Speech transcriptions also contain truncated words, orthographic reductions, etc. Normalizing or rewriting such texts using ordinary words is an important issue for various applications. There is a greater need for work on text normalization, as it forms an important component of all areas of language and speech technology. 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. Considering the above, as well as the context of genericity, producing reusable components for language-and-task-specific development is an important goal. This section describes SPPAS text normalization and concentrates on the aspects of methodology and linguistic engineering which serve to develop this multi-purpose multi-lingual text corpus normalization method.

SPPAS implements a generic approach, i.e. 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 and rapid adaptation of other language dependent modules. In the same way, for a new task, a module can be inherited from general processing modules, and adapted rapidly to create other specific modules.

The first step is to determine which modules to use, some are shared (the modules which do not depend on the language), and some are variable modules (language-dependent modules). This splitting and specification of work is really important. For modeling a new language, the shared modules will be inherited and the variable modules will be adapted to that language. It will economize the time needed to complete corpus normalization. The key idea is to concentrate the language knowledge in a set of lexicons and to develop modules which implement rules to deal with the knowledge elements. Shared modules are listed below:

- *Basic unit splitting module:* a segmentation module based on white spaces for Romanized languages and character-based for the other languages.
- *Replacing module:* implements a dictionary look-up algorithm to replace a string by another one. It is mainly used to replace special symbols like ° (degrees), for example.
- *Lowerize module:* used to convert the character-case.
- *Word-tokenization module:* fixes a set of rules to segment strings including punctuation marks for Romanized languages. This algorithm splits strings into words on the basis of a dictionary and a set of manually established rules. For example, in French “trompe-l'oeil” (*sham*) is an entry in the vocabulary and it will not be segmented. On the other hand, an entry like “l'oeil” (*the eye*) occurring in another context will be segmented into two separate words.
- *Sticking module* implements an algorithm to concatenate strings (or characters) into words based on a dictionary with an optimization criteria: *longest matching*.
- *Removing module* can be applied to remove strings of a text. The list of strings to remove is defined in a separate file. For certain applications, it is relevant for example to remove punctuation marks.

Apart from the abovementioned shared modules, SPPAS also includes several language-specific modules. One of them is the optional *number to letter module*. For example, the number “123” is normalized as “one_hundred_twenty-three” for English and “ciento_veintitres” in Spanish. It is thus

necessary to implement this module for each new language if numbers are used in the orthographic transcription. Adding a new language only consists of adding the list of tokens in the appropriate directory of the SPPAS package, and eventually writing the number to letter conversion. It means also that any phonetician can edit/modify the lexicon to get the expected result.

Another specific module has been developped to deal with enriched orthographic transcriptions. From the manual EOT (Enriched Orthographic Transcription), two types of transcriptions are automatically derived by the tokenizer: the "standard transcription" (a list of orthographic tokens/words) and the "faked transcription" that is a specific transcription from which the obtained phonetic tokens are used by the phonetization system. The following example illustrates an utterance text normalization extracted from the CID corpus in French:

> **Transcription:** j'ai on a j'ai p- (en)fin j'ai trouvé l(e) meilleur moyen c'était d(e) [loger,locher] chez des amis (*I've we've I've - well I found the best way was to live in friends' apartment'*)
> **Standard transcription:** j' ai on a j' ai p- enfin j' ai trouvé le meilleur moyen c'était de loger chez des amis
> **Faked transcription:** j' ai on a j' ai p- fin j' ai trouvé l meilleur moyen c'était d locher chez des amis

The standard one is "human-readable" and can be used for further processing by any automatic system, e.g., an automatic syntax analysis. The faked one is useful mainly for the grapheme-to-phoneme conversion system. In the case of standard orthographic transcription instead of EOT, both the generated standard and faked transcriptions are identical. See Bigi et al. (2012) for an evaluation of the impact of such EOT on the automatic phonetization system of SPPAS.

We applied the SPPAS automatic tokenizer on the 16 files of the French CID corpus, which were fully transcribed with EOT. Each file represented the transcription of one hour of speech in the context of eight dialogues. This process was accomplished in 95sec with SPPAS version 1.7.2 on a 2009-Desktop PC. The result was a set of 16 files containing the normalized text (a total of 120,000 tokens) including standard and faked transcriptions.

3.2 Phonetization

Phonetic transcription of text is an indispensable component of text-to-speech systems and is used in acoustic modeling for automatic speech recognition and other natural language processing applications. Generally, grapheme-to-phoneme conversion is a complex task, for which a number of diverse solutions have been proposed. It is a structure prediction task; since both the input and output are structured, consisting of sequences of letters and phonemes, respectively. It can be implemented in many ways, often roughly classified into dictionary-based and rule-based strategies, although many intermediate solutions exist. In the context of our study, the phonetization process takes the normalized transcription of the speech signal as input and produces the supposed pronunciation. The phonetization of speech corpora requires a sequence of processing steps and resources in order to convert the normalized text into its constituent phones.

SPPAS implements a dictionary-based approach, which is relatively language-independent. The dictionary includes phonetic variants that are proposed for the aligner to choose the phoneme string. The hypothesis is that the answer to the phonetization question can be found in the speech signal. Consequently, an important step is to build the pronunciation dictionary, where each word in the

vocabulary is expanded into its constituent phones, including pronunciation variants. Depending on the language, the availability of such resources varies. In the SPPAS data set, the dictionary includes a large set of entries for English, French, Italian, Polish, an acceptable number of entries for Catalan, Mandarin Chinese, Spanish, Japanese, Cantonese, and a rather poor number of entries for Taiwan Southern Min. In addition, SPPAS implements an algorithm for phonetization of unknown words (e.g., proper names, speech reductions or mispronunciations). The present grapheme-to-phoneme conversion system is based on the idea that given enough examples it should be possible to predict the pronunciation of unseen words purely by analogy. The system is then applied to missing words during the phonetization process (and not during a training stage), and is only based on knowledge provided by the dictionary. The algorithm consists of exploring the unknown entry from left to right, then right to left, to find the longest strings in the dictionary. Since SPPAS-Phonetization only uses the pronunciation dictionary either for known or unknown words, the quality of such an annotation depends mainly on the quality of a particular resource. Another consequence of such a system is that adding a new language in SPPAS-Phonetization only consists in adding the pronunciation dictionary in the appropriate directory of the SPPAS package. It also means that any phonetician can use their own dictionary.

We applied the SPPAS automatic phonetizer on the 16 normalized files of the French CID corpus. The process was accomplished in 71sec with SPPAS version 1.7.2 on a 2009-Desktop PC. The result was a set of 16 files containing the phonetized transcription, including pronunciation variants.

3.3 Speech segmentation

Phoneme alignment is the task of proper positioning of a sequence of phonemes in relation to a corresponding continuous speech signal. In the alignment task, we are given a speech utterance along with the given phonetic representation for that utterance. Our goal is to generate an alignment between the speech signal and the phonetic representation. Manual alignment has been reported to take between 11 and 30 seconds per phoneme (Leung and Zue, 1984). An automatic time-alignment system is then essential for the annotation of large corpora.

SPPAS is based on the use of the Julius Speech Recognition Engine (Lee et al., 2001). 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.

The Julius alignment task processes in two-steps: The first step selects the phonetization and the second step performs the segmentation. A finite state grammar that describes sentence patterns to be recognized and an acoustic model are needed. This grammar essentially defines constraints on what the Speech Recognition Engine can expect as input. SPPAS generates the grammar automatically from the phonetized files. Speech alignment also requires an acoustic model in order to align speech. This involves a file that contains statistical representations of each of the distinct sounds in a language. The original Julius distribution only includes Japanese acoustic models. However since it can use acoustic models of HTK-ASCII format (a common format used by many systems), this system can also be adapted to other languages. Consequently, any user can train it's own acoustic model, or get it from the web, and integrate it in SPPAS.

Most of the acoustic models already included in SPPAS were trained by the author of this paper with HTK by taking a training corpus of speech, previously segmented into utterances and phonetized. Ideally, the phones would have unique articulatory and acoustic correlates. But acoustic properties of a given phone can depend on the phonetic environment. These co-articulation phenomena motivated the adoption of context-dependent models such as triphones, for each language we had enough data for training. To train such acoustic models, the training procedure is based on the VoxForge tutorial[1], except that VoxForge suggests using only word transcription as input, and we allow (and prefer) to use phonetized ones. The outcome of this training procedure is dependent on the availability of accurately annotated data and on good initialization. Acoustic models were trained from 16 bits, 16000 Hz wav files. This procedure had three main steps:

- data preparation,
- monophones generation,
- triphones generation.

Step 1 establishes the list of phonemes, plus silence and short pauses. It converts the input data (phonetization of the corpus) into an HTK-specific data format. It codes the (audio) data in a process known as "parameterizing the raw speech waveforms into sequences of feature vectors". Step 2 involves monophones generation. It creates a Flat Start Monophones model by defining a prototype model and copying this model for each phoneme. Then, this flat model is re-estimated using the provided data files to create a new model. Step 3 creates tied-state triphones. From our previous studies on French and Italian, we observed that five minutes of manually-time-aligned data are sufficient to train the initial model; and we found that about 10-30 minutes of manually-phonetized data are required to train a good monophone model. The orthographic transcription of several hours of speech will allow one to train a triphone model. As a consequence, any phonetician who had already created such a corpus for any language could share it privately with the author of SPPAS for a new acoustic model to be trained and publicly shared with the community.

We applied the SPPAS automatic aligner on the 16 audio files of the CID corpus, which were already converted to wav/mono/16000Hz/16bits, as the default in SPPAS. The process of time-aligning these 14000 IPUs was accomplished in 84min with SPPAS version 1.7.2 on a 2009-Desktop PC. The result was a set of 16 files containing the time-aligned phonemes and tokens (as shown in tiers 2, 3 and 4 of Figure 1).

3.4 Syllabification

The syllabification implemented in SPPAS is a rule-based system based on time-aligned phonemes. This phoneme-to-syllable segmentation system is based on two main principles:

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

These two principles focus on the problem of finding a syllabic boundary between two vowels. Phonemes were grouped into classes and rules established to deal with these classes. We defined general rules as well as a small number of exceptions. Consequently, the identification of relevant classes is important for such a system. The rules follow usual phonological statements for most of the corpora and Romance

[1] http://www.voxforge.org

languages. An external configuration file indicates phonemes, classes and rules. This file can be edited and modified by any user to adapt the syllabification to any language or phoneme encoding. In the current version of SPPAS the respective sets of rules are available for French, Italian and Polish.

3.5 Self- and Other-repetitions

Other-repetition (OR) is a device involving the reproduction by a speaker of what another speaker has just said. Other-repetition has been found as a particularly useful mechanism in face-to-face conversation due to the presence of discursive or communicative functions. Among their various functions in discourse, repetitions serve the purpose of facilitating comprehension by providing less complicated discourse, while also establishing connection between various stages of discourse (cohesion), and also function as a device for getting or keeping the floor. SPPAS implements a semi-automatic method to retrieve other-repetition occurrences (Bigi et al. 2014). A key-point is that the proposed automatic detection is based on observable cues which can be useful for OR's identification from the time-aligned tokens. SPPAS captures repetitions which can be an exact repetition (named strict echo) or a repetition with variation (named non-strict echo). The rules of this system have been adapted to the detection of self-repetitions in the context of a study presented in (Tellier et al. 2012). As such, this method is intrinsically language-independent.

4 Conclusion

This paper described the automatic annotation systems included in SPPAS, a computer software tool designed and developed by the author to handle multiple language corpora and/or tasks with the same algorithms in the same software environment. Only the resources (e.g., dictionaries, lexicons, acoustic models) are language-specific and the approach is based on the simplest resources possible. The present work emphasizes new practices in the methodology of tool developments: considering the problems with a generic multi-lingual aspect, sharing resources, and putting the end-users in control of their own computing.

We hope this work will be helpful for the linguistic research community, and especially for those involved in speech research, as far as possible. Phoneticians are of crucial importance for resource development as they can contribute to improve the resources used by automatic systems. In the case of SPPAS, the improved software 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, free to share the software and resources with other people, free to modify the software and resources, and free to publish their modified versions of the software and resources.

References

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

Alazard, C.,C. Astésano and M. Billières (2012). MULTIPHONIA: a MULTImodal database of PHONetics teaching methods in classroom InterActions. Language Resources and Evaluation Conference, Istanbul (Turkey).

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman (2001). Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication 33, no. 1: 5-22.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. Traitement automatique des langues 49, no. 3: 1-30.

Blache, P., R. Bertrand, B. Bigi, E. Bruno, E. Cela, R. Espesser, G. Ferré, M. Guardiola, D. Hirst, E.-P. Magro, J-C Martin, C. Meunier, M-A. Morel, E. Murisasco, I. Nesterenko, P. Nocera, B. Pallaud, L. Prévot, B. Priego-Valverde, J. Seinturier, N. Tan, M. Tellier, S. Rauzy (2010). Multimodal Annotation of Conversational Data. The Fourth Linguistic Annotation Workshop, ACL 2010, pages 186-191, Uppsala, Suède.

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

Bigi, B. (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, B., R. Bertrand and M. 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, B., T. Watanabe and L. 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.

Bigi, B., K. Klessa, L. Georgeton and C. Meunier (2015). A syllable-based analysis of speech temporal organization: a comparison between speaking styles in dysarthric and healthy populations. INTERSPEECH, Dresden (Germany).

Bigi, B. and J. Saubesty (2015). Searching and retrieving multi-levels annotated data. Gesture and Speech in Interaction. Nantes (France).

Bigi, B. and R. Fung (2015). Automatic Word Segmentation for Spoken Cantonese. Oriental Chapter of International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques. Shanghai (China).

Boersma, P., and D. Weenink (2001). Praat, a system for doing phonetics by computer. pages 341-345.

Chiarcos, C., S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede (2008). A flexible framework for integrating annotations from different tools and tagsets. Traitement Automatique des Langues 49, no. 2: 271-293.

Gibbon, D. (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.

Gorisch, J., C. Astésano, E. Gurman Bard, B. Bigi, and L. 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.

Herment, S., A. Tortel, B. Bigi, D. Hirst, A. Loukina (2014). AixOx, a multi-layered learner's corpus: automatic annotation. Specialisation and Variation in Language Corpora. Linguistic Insights: Studies in Language and Communication. Eds Ana Díaz-Negrillo and Francisco Javier Díaz-Pérez. pages 41-76, vol. 179. ISBN: 978-3-0343-1316-2.

Hirst, D.J. and R. 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, D.J. (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.

Klessa, K., M. Karpiński, and A. Wagner (2013). Annotation Pro-a new software tool for annotation of linguistic and paralinguistic features. In Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, pp. 51-54.

Klessa, K. (2015). Annotation Pro [Software tool]. Version 2.2.6.0. Retrieved from: http://annotationpro.org/ on 2015-05-19.

Lamere, P., P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf (2003). The CMU SPHINX-4 speech recognition system. In IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, vol. 1, pp. 2-5. 2003.

Lee, A., T. Kawahara and K. Shikano (2001). Julius --- an open source real-time large vocabulary recognition engine. In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691-1694.

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

Leung, H.C., and V.W. Zue (1984). A procedure for automatic alignment of phonetic transcriptions with continuous speech. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84., vol. 9, pp. 73-76. IEEE.

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

Stallman R. (2002). Free Software, Free Society: Selected Essays of Richard M. Stallman. Retrieved on 2015-09-27 from: https://www.gnu.org/philosophy/fsfs/rms-essays.pdf

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

Teston, B., A. Ghio, and B. Galindo (1999). A multisensor data acquisition and processing system for speech production investigation. In International Congress of Phonetic Sciences (ICPhS), pp. 2251-2254. University of California.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes (2006).ELAN: a professional framework for multimodality research. In Proceedings of LREC.

Young, S. J., and Sj Young (1993). The HTK hidden Markov model toolkit: Design and philosophy. University of Cambridge, Department of Engineering.

Yu, J. (2013). Timing analysis with the help of SPPAS and TGA tools. Tools and Resources for the Analysis of Speech Prosody, Aix-en-Provence, France. Eds B. Bigi and D. Hirst, ISBN: 978-2-7466-6443-2.