Orthographic Transcription: which Enrichment is required for phonetization?

Brigitte Bigi, Pauline Péri, Roxane Bertrand

Laboratoire Parole et Langage
CNRS & Aix-Marseille Université,
5, avenue Pasteur, BP80975
13604 Aix-en-Provence France
brigitte.bigi@lpl-aix.fr

Abstract

This paper addresses the problem of the enrichment of transcriptions in the perspective of an automatic phonetization. Phonetization is the process of representing sounds with phonetic signs. There are two general ways to construct a phonetization process: rule based systems (with rules based on inference approaches or proposed by expert linguists) and dictionary based solutions which consist in storing a maximum of phonological knowledge in a lexicon. In both cases, phonetization is based on a manual transcription. Such a transcription is established on the basis of conventions that can differ depending on their working out context. This present study focuses on three different enrichments of such a transcription. Evaluations compare phonetizations obtained from automatic systems to a reference phonetized manually. The test corpus is made of three types of speech in French: conversational speech, read speech and political debate. A specific algorithm for the rule-based system is proposed to deal with enrichments. The final system obtained a phonetization of about 95.2% correct (from 3.7% to 5.6% error rates depending on the corpus).
Keywords: transcription, speech, phonetization

1. Introduction

The study presented in this paper is part of the OTIM project (also called TOMA - Tools for Multimodal Information Processing)[1] described in (Blache et al., 2009). The project focuses on the different requirements and needs in the perspective of multimodal annotations. A broad-coverage approach, aiming at annotating a large set of linguistic domains is proposed. The multimodal annotation is faced with the necessity of encoding different information types, from different domains, with different levels of granularity. OTIM aims to develop such a multimodal annotation scheme and tools for face to face interaction. This implies technical and methodological levels to produce high quality multimodal annotations (Blache et al., 2010).
In this field, transcription of the speech signal is the first annotation. Depending on the focus of a study, a transcription can be annotated following various conventions. Even more so, in a multimodal perspective, the transcription has to satisfy the needs and constraints of each domain. The difficulty comes from the fact that each domain investigated a different perspective and had different objectives; researchers interested in morpho-syntax level or in phonetic or prosodic level have not the same needs.
Phonetic level is one of the domains annotated in the OTIM project. Phonetization is the process of representing sounds with phonetic signs. There are two general ways to construct a phonetization process: dictionary based solutions which consist in storing a maximum of phonological knowledge in a lexicon and rule based systems with rules based on inference approaches or proposed by expert linguists. In both cases, phonetization is based on the manual transcription. When a speech corpus is transcribed into a written text, the transcriber is immediately confronted with the following question: how to reflect the reality of oral speech in a corpus? Conventions are then designed to provide a set of rules for writing speech corpora. These conventions establish which phenomena have to be annotated and also how to annotate them.

Numerous studies have been carried otu in prepared speech, as for example for broadcast news (ESTER2, 2008). However, conversational speech refers to an activity more informal, in which participants have constantly to manage and negociate turn-taking, topic (among other things) "on line" without any preparation. As a consequence, numerous phenomena appear such as hesitations, repeats, feedback, backchannels, etc. Other phonetic phenomena such as non-standard elision, reduction phenomena (Meunier and Espesser, 2011), truncated words, and more generally, non-standard pronunciations are also very frequent. All these phenomena can impact on the phonetization.

This paper focuses on phenomena that are mentioned in the transcription and the consequence of this annotation on the quality of the phonetization. The aim was to compare some phonetization approaches based on various transcription enrichments and to answer the question: which speech phenomena are needed to be transcribed to obtain a good phonetization?

Section 2. reports three different transcription enrichments, and a description of the corpus used in this study. Sections 3. presents a dictionary-based approach for phonetization (language independent approach), and a rule-based system dedicated to the phonetization of French. This latter system was initially developed to deal with a standard transcription. Section 4. reports a tree-based algorithm that adapts to transcription enrichments. Finally, experiments are reported in Section 5. Evaluation were carried out by comparing the automatic phonetization systems to a manual reference. The hand-made test corpus represents about 7 minutes of speech and is divided into three types of speech: conversational data, read speech and a political debate.

[1] http://www.lpl-aix.fr/∼otim/

2. Enriched Orthographic Transcription

2.1. Transcription conventions

The transcription process follows specific conventions. The result is what is called an enriched orthographic construction. In this study, three enrichments were selected.
The first one represents the text as a standard orthographic written text. For example, if the speech signal includes specific productions like reductions, the transcription must contain the expended written text: for example *je suis* pronounced as [ʃɥi] instead of the standard pronunciation [ʒsɥi], and the same for *il y a* that can be pronounced [iɑ]. Other specific speech phenomena are also ignored.
In the second transcription, transcribers provided an enriched orthographic transcription, which included, for example, manually annotating non-standard events such as: truncated words, laughter, etc. Compared to the previous one, this enriched orthographic transcription includes the following specific speech phenomena:

- short pauses, annotated '+'
- various noises, annotated '*'
- laughter, annotated '@'
- filled pauses, annotated 'euh'
- truncated words, annotated with a '-'
- repeats

Moreover, the transcription was not systematically expanded to the written text: the speech sound [iɑ] was transcribed as *y a*. But specific reductions like [ʃɥi] were transcribed as a standard orthographic written text *je suis*.
The third transcription used in this study represented both transcriptions: the orthographic written text (as the previous convention) and, if any, the specific production using an orthographic written text the nearest as possible of what the transcriber could hear. Thereby, from this manual transcription, two derived transcriptions can be generated automatically: the "real orthographic" transcription (the list of orthographic tokens) and a specific transcription from which the obtained phonetic tokens are used by the phonetization system. This latter spelling is called "faked spelling" in this paper. If a token was not modified manually by the transcriber, it is then supposed to be pronounced in a standard way. This is the approach proposed in the OTIM project. These two versions of transcription, synchronized and aligned on the signal, are used either by the morpho syntactic and discourse level, or by the phonetic and prosodic level.
Specific productions have a direct consequence on the phonetization procedure:

- Elision is the omission of one or more sounds (such as a vowel, a consonant, or a whole syllable). Non-standards elisions are explicit in this transcription, manually annotated by parenthesis of the omitted sounds. For example:

  `j'ai on a j'ai p- (en)fin j'ai trouvé l(e) meilleur moyen c'(é)tait d(e) loger chez des amis`
  'I've we've I've - well I found the best way was to live in friends' apartment'

  Consequently, the phonetizer will not produce phonemes for elision in the words *enfin*, *le*, etc. Another word frequently produced with elision is *parce que* phonetized as /pɑsk/ or even /psk/ instead of /pɑʁsk/.

- Transcribers also mentioned particular phonetic realizations by using brackets, such as the pronunciation of specific words, pronounced schwa, etc. For example:

  `[elle, è] dormait`
  'She slept'

  `du [movetrack, mouvtrac] ouais de de l'[EMA, euma]`
  'of movetrack yeah of of EMA'

  `faire des [stats, stateu]`
  'to do stats'

- Optional liaisons were also manually mentioned in this enriched transcription.

2.2. Test corpus description

To our knowledge, there is no publicly available corpus phonetically transcribed for French that could be used for this study. We thus constructed such a corpus. This annotation was performed by a phonetician, well skilled in the perception and transcription of speech sounds. Phonetic transcription is therefore a very time consuming task.
In parallel, the corpus was transcribed using the three transcription enrichments described previously in this paper.
The test corpus was based on parts of three different French corpora downloaded from the SLDR - Speech & Language Data Repository:

http://www.sldr.org

About two minutes of each corpora (about 7 minutes altogether) were manually segmented and transcribed.
The first one was extracted from the corpus created during the "Amennpro" project, a Franco-British partnership program that attempts to develop methods of automated evaluation of rhythm in non native speech. Only French speech as spoken by French native speakers were selected. This audio corpora was called AixOx (Herment et al., 2012). This corpus is related to "read speech", as speakers were asked to read paragraphs made of about 3 to 6 sentences. The second part of the test corpus was extracted from CID - Corpus of Interactional Data (Bertrand et al., 2008). CID is an audio-video recording of 8 hours of spontaneous French dialogues, 1 hour of recording per session. Each dialogue involved two participants of the same gender. One of the following two topics of conversation was suggested to participants: conflicts in their professional environment or funny situations in which participants may have found themselves. Finally, the test corpus contained an extract of a political debate; this corpus was named Grenelle (Bigi et

al., 2011). Grenelle concerns a political debate on environmental issue recorded at the French National Assembly on the 4th of May 2010. While AixOx and CID have been recorded in a sound attenuated room, Grenelle has been recorded in a naturalistic environment.

	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

Table 1: Test corpus description

Table 1 reports a detailed description of the test corpus. It is important to mention that the three corpora were equally represented (at least for the present study): about the same duration and the same number of phonemes. This corpus represented the expected phenomena like silent pauses or truncated words. Moreover it contained the expected rate of these phenomena depending on the speech types. Particularly, conversational data (CID) compared to both other speech types included a very high rate of hesitations. CID also contained laughter but not in the other corpora; and CID contained a larger number of elisions and special pronunciations. The AixOx corpus that represented read speech was made up a larger set of special pronunciation as it was not expected in this type of speech. This was mainly due to a regional accent of one speaker in the test set.
Two examples of each corpus are presented in Tables 3, 4 and 5 depending on the transcription annotation (an English translation of these examples is proposed in Table 2). Table 3 is related to the standard orthographic transcription. In case of read speech, this transcription corresponds to the script that participants had to read.

3. Dictionary-based phonetization

Clearly, there are different ways to pronounce the same utterance. Different speakers have different accents and tend to speak at different rates. A system based on a dictionary solution consists in storing a maximum of phonological knowledge in a lexicon. Phonetic variants are proposed to an aligner to choose the phoneme string. By using this approach, the hypothesis is that the answer to the phonetization question is in the signal. This approach can take as input a standard orthographic transcription and some enrichments only if the acoustic model includes them.
Experiments reported in this paper were carried out using SPPAS - SPeech Phonetization Alignment and Syllabification (Bigi and Hirst, 2012). SPPAS is a tool to produce automatically annotations which includes utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. The whole procedure is a succession of 4 automatic steps. Resulting alignments are a set of TextGrid files. TextGrid is the native file format of the Praat software which became the most common tool for phoneticians (Boersma and Weenink, 2009). It is currently implemented for French, English, Italian and Chinese and there is a very simple procedure to add other languages. An important point for software which is intended to be widely distributed is its licensing conditions. SPPAS uses only resources, tools and scripts which can be distributed under the terms of the GPL license. SPPAS tools and resources are freely available at the URL:

http://www.lpl-aix.fr/∼bigi/sppas/

To perform the phonetization, an important step is to build the pronunciation dictionary, where each word in the vocabulary is expanded into its constituent phones. The phonetization is the equivalent of a sequence of dictionary lookups. This approach supposes that all words of the speech transcription are mentioned in the pronunciation dictionary otherwise a pronunciation is constructed from inputs of the dictionary using a longuest-matching algorithm. Actually, some words can correspond to several entries in the dictionary with various pronunciations. Thus, the dictionary contains a set of possible pronunciations of each words, including accents, reduction phenomena and liaisons like “je suis”:

- /ʒsɥi/ is the standard pronunciation,
- /ʒsɥiz/ is the standard pronunciation plus a liaison,
- /ʒəsɥi/ is the South of France pronunciation,
- /ʒəsɥiz/ is the South of France pronunciation plus a liaison,
- /ʃɥi/ is a frequent specific realization.

The French dictionary included in SPPAS contains 350k entries and 300k variants. SPPAS determines the pronunciation during the alignment (also called phonetic segmentation) step because the pronunciation generally can be observed in the speech.
Phonetic segmentation is the process of aligning speech with its corresponding transcription at the phone level. The alignment problem consists in a time-matching process between a given speech utterance and a phonetic representation of the utterance. The goal is to generate an alignment between the speech signal and its phonetic representation. SPPAS is based on the Julius Speech Recognition Engine (SRE). To perform alignment, a finite state grammar that describes sentence patterns to be recognized and an acoustic model are needed. A grammar essentially defines constraints on what the SRE can expect as input. This is a list of words; and each word has a set of associated list of phonemes, extracted from the dictionary. When given a speech input, Julius searches for the most likely word sequence under constraint of the given grammar.
Speech Alignment also requires an Acoustic Model in order to align speech. An acoustic model is a file that contains statistical representations of each of the distinct sounds of

one language. Each phoneme is represented by one of these statistical representations. These are called Hidden Markov models (HMMs).
The French acoustic model included in SPPAS (version 1.4) was trained using HTK (Young and Young, 1994) from 7h30 of conversational speech extracted from CID and 30 minutes of read speech extracted from AixOx (previously segmented in utterances and automatically phonetized).

4. Rule-based phonetization

4.1. Basic phonetization system

The phonetization system used in this study was LIA_Phon (Bechet, 2001) which is distributed under the term of the GPL license. LIA_Phon contains a set of scripts that transform a raw text into its phonetic form. There are 3 main steps in this process:

1. Formatting the text,
2. POS tagging + accentuation,
3. Grapheme-to-Phoneme transcription.

The POS-tagger aims at the pronunciation disambiguation, for example:

- est 'is' (verb) is pronounced /e/
- est 'east' (noun) is pronounced /ɛst/

POS tagging was performed using the LIA_TAGG. LIA_TAGG contains a set of scripts in order to clean, format, tagg and bracket French or English texts. It is based on a set of 103 tags using short-cuts like: NMS for the tuple (name, masculine, singular), ADV for adverbs, V3S for the tuple (verb, 3rd person, singular), etc. LIA_Phon and LIA_TAGG tools and resources are freely available at the URL:

http://pageperso.lif.univ-mrs.fr/∼frederic.bechet/

LIA_Phon was conceived to take as input a standard orthographic transcription. The pronunciation was supposed to correspond to a standard French. To deal with the two enrichments of transcriptions (Section 2.), the faked spelling has been sent to the phonetizer. But faked entries were recognized as unknown words and the tagger still had to assign a tag. This necessarily implies tag errors. Thus this could cause phonetization errors, not only on the concerned entry but also on the n previous or following entries due to the use of n-gram models in these tools.
In the following example, the use of the LIA_Phon with a faked spelling produced one phonetization error on the word *"dit"* which was pronounced [d]. The automatic phonetization was [de], because the faked-word was *"d"* and it was recognized as a noun and then it was spelled.
Example with a real spelling is: `oui ben oui puisque de toute façon il m'a dit il a trouvé un appart et tout là-haut donc c'est que euh`[2]

[2]yeah so anyway because he said he found an apartment and all thereby thus it's that hum

Example with a faked spelling is: `oui bè oui pusque tfaçon i m'a d il a trouvé un appart et tout là-haut donc c'est qu euh`
Table 6 illustrates tag errors that are produced by the use of a faked orthograph directly in the LIA_TAGG. For this sentence (extracted from CID corpus), about 32% of entries obtained a wrong tag.

Real Ortho.	POS-Tag	Faked Ortho.	POS-Tag	Tag Error
oui	ADV	oui	ADV	
ben	ADV	bè	AFS	X
oui	ADV	oui	ADV	
puisque	COSUB	pusque	AFS	X
de	PREPADE			X
toute	DETFS			X
façon	NFS	tfasson	NFS	
il	PPER3MS	i	NMS	X
m'	PPOBJMS	m'	PPOBJMS	
a	VA3S	a	VA3S	
dit	VPPMS	d	NMS	X
il	PPER3MS	il	PPER3MS	
a	VA3S	a	V3S	
trouvé	VPPMS	trouvé	VPPMS	
un	DETMS	un	DETMS	
appart	NMS	appart	NMS	
et	COCO	et	COCO	
tout	ADV	tout	ADV	
là-haut	ADV	là-haut	ADV	
donc	COCO	donc	COCO	
c'	PPER3MS	c'	PPER3MS	
est	VE3S	est	VE3S	
que	COSUB	qu	AMS	X
euh	ADV	euh	ADV	

Table 6: LIA_TAGG outputs depending on the spelling given as input

A suitable adaptation of such a tool to deal with enriched orthographic transcriptions is proposed in this paper.

4.2. Tree-based phonetization system

Initially, the phonetization system dealt only with standard orthographic transcriptions. The system could be used with some enrichments like repeats or truncated words because it included a French-specific algorithm for the phonetization of unknown words. The phonetization process was based on the use of a POS-tagger and these phenomena could cause errors.
A tree-based approach is proposed. It consists in sending the real orthographic transcription to the tagger to obtain good tags. Then, the tuple containing the faked spelling plus the tags were sent to the phonetizer. Figure 1 illustrates 2 examples of the use of this algorithm. Gray circles represent nodes of the tree. Nodes can be of types: root, token, laugh, pronunciation, elision, liaison, trunc or pause. Elision and pronunciations could have two children: left-child was the real orthographic written text and right-child corresponded to the faked spelling. First, the tree was explored to phonetize automatically (and independently) each

leaf of type: pause, laugh, trunc and liaison. Then, the tree was explored by using only the left part of each node, and by ignoring pauses, laugh, trunc and liaisons. This sentence was sent to the LIA_TAGG to obtain the POS-tags (orange-color in the examples). These POS-tags were then copied to right leaves for trunc and pronunciations nodes. The tree was then explored to get the right part of each node and the associated POS-tag (also by ignoring pauses, laugh, trunc and liaisons). Lastly, the tree was explored to obtain the phonetization.
This algorithm was implemented in *ESPPAS - Enriched-SPPAS*, a plugin to *SPPAS* also available under the terms of the GPL license (for unix-based systems only).

5. Results

The most common and direct form of evaluation is comparing the automatic phonetization to a manual one. Evaluations were performed using Sclite. Sclite is provided by the National Institute of Standards and Technology (Sclite, 2009). Accuracy is calculated as a function of phonemes, by estimating the sum of the following errors:

- substitution (Sub), example : ɛ̃ / ɔ̃
- deletion (Del), example : p ə t i / p t i
- insertion (Ins), example : ʒ / ʒ ə

Evaluations considered a reduced set of phonemes by combining the following pairs: o/ɔ, e/ɛ, ’/i. These 3 cases was related to about 2.7% of substitution errors, independently on the corpus or the transcription.
For the measurement of accuracy rates, the manual phoneme transcription of the test files was compared to:

- the output that *SPPAS (version 1.4)* system produced (dictionary-based approach),
- the output that *LIA_Phon* system produced (rule-based approach),
- the output that *Enriched-SPPAS* system produced (rule-based approach, with a tree-based algorithm).

	System	Sub	Del	Ins	Err
AixOx					
Standard Trs	Lia_Phon	1.4	5.0	3.0	9.5
Standard Trs	SPPAS 1.4	3.6	4.5	2.8	10.8
Grenelle					
Standard Trs	Lia_Phon	1.1	2.8	4.1	8.0
Standard Trs	SPPAS 1.4	2.3	2.8	3.8	8.8
CID					
Standard Trs	Lia_Phon	2.8	4.5	10.0	17.3
Standard Trs	SPPAS 1.4	3.6	4.9	6.0	14.5

Table 7: Phonetization errors (in %), obtained from a standard transcription

Results using the standard orthographic transcription are presented in Table 7. For both systems, the automatic phonetization is very different from what it is expected. This is the case independently of the corpus, but significantly for CID using LIA_Phon. A detailed analysis shows that major part of error are related to insertions, specially for CID that contained a large set of specific pronunciations due to the type of speech. Deletion errors are also observed because this transcription did not include specific speech phenomena that was not automatically phonetized but that the manual reference included.

	System	Sub	Del	Ins	Err
AixOx					
Enriched1 Trs	Lia_Phon	1.4	2.3	2.9	6.5
Enriched1 Trs	SPPAS 1.4	3.1	2.3	2.8	8.2
Grenelle					
Enrichie1 Trs	Lia_Phon	1.0	1.2	4.1	6.3
Enriched1 Trs	SPPAS 1.4	1.7	1.7	3.9	7.3
CID					
Enriched1 Trs	Lia_Phon	2.7	1.4	10.3	14.4
Enriched1 Trs	SPPAS 1.4	3.3	2.3	6.9	12.5

Table 8: Phonetization errors (in %), obtained with the first enriched transcription

Table 8 presents results by transcribing with the first enrichment. It included a set of speech phenomena (silent pauses, filled pauses, repeats, etc.) but not specific realizations. This enrichment allowed automatic systems to produce a significantly better phonetization. The LIA_Phon improved its scores of about 3.0% for CID and AixOx and 1.7% for Grenelle. SPPAS is the better system to deal with conversational speech. The use of the LIA_Phon in a tree-based approach produced the same scores as the use of the LIA_Phon directly. The enrichments proposed in the transcription was particularly interesting to reduce deletion errors. However, a large set of deletion errors still occurred in the AixOx phonetization due to the regional accent of one speaker in the test corpus. This speaker added schwas that are not commons in standard French and he pronounced standard elisions. Despite the improvements this enrichment can provide, both automatic phonetization systems produced a large set of errors: compared to the previous one, there was no consequence on substitutions or insertions.

	Algorithm	Sub	Del	Ins	Err
AixOx					
Enriched2 Trs	LIA_Phon	1.3	1.8	2.5	5.6
Enriched2 Trs	ESPPAS	1.4	1.4	2.4	5.2
Enriched2 Trs	SPPAS 1.4	3.0	2.1	3.1	8.2
Grenelle					
Enriched2 Trs	LIA_Phon	1.3	1.0	1.7	4.0
Enriched2 Trs	ESPPAS	1.2	1.2	1.4	3.7
Enriched2 Trs	SPPAS 1.4	2.1	1.7	2.4	6.2
CID					
Enriched2 Trs	LIA_Phon	1.8	1.3	3.4	6.5
Enriched2 Trs	ESPPAS	1.7	1.3	2.6	5.6
Enriched2 Trs	SPPAS 1.4	2.4	2.7	4.5	9.5

Table 9: Phonetization errors (in %), obtained with the second enriched transcription

Table 9 presents results using the second enrichment. The basic idea of this enrichment was to suppose that the pronunciation was standard, except if the manual transcription mentioned something else. The transcriber disambiguated pronunciations. Thus, this enrichment was particularly adapted to automatic rule-based systems which, by default, proposed a standard pronunciation.

Even by using this enrichment, the dictionary-based approach did not suppose any kind of pronunciations (except for enriched-entries) and proposed phonetic variants like for the previous enrichments. This is the reason of 1/ the small improvements compared to the previous enrichment and 2/ the lower performances of SPPAS in this experiment compared to the other systems. However, SPPAS used an acoustic model trained on 8h of speech and better results could be potentially expected by using an acoustic model trained from a larger corpus. Unlike rule-based systems, it is also important to note that the algorithms of a dictionary-based approach, as implemented in SPPAS, are completely language-independent (only resources are language-dependent).

Performances of the LIA_Phon were significantly improved by using this enrichment, specially for CID where the number of errors was divided by 2. This enrichment allowed the system to reduce insertions errors: divided by 3 for CID and divided by 2.5 for Grenelle. This enrichment also allowed to reduce deletions errors for AixOx (particularly for the speaker with a regional accent). Despite POS-tags errors (see Section 4.), this phonetization was quitey good, and there was just a little room for improvements.

Performances of the tree-based algorithm (also based on the use of LIA_Phon) improved performances of about 7% relative gain for AixOx and Grenelle and about 14% relative gain for CID. Errors that were corrected were mainly concerning insertions. This was essentially due to POS-tag errors that induced the phonetizer to spell words instead of phonetizing them.

6. Conclusion

This paper examined three transcription enrichments and showed how it impacted on the performances of automatic phonetization. A dictionary-based system (SPPAS) and a rule-based system (LIA_Phon) were compared. Evaluations were carried out by using a French test corpus manually phonetized by an expert, related to three different types of speech: conversational speech, read speech, political debate. Results indicated clearly that the richer transcription the better phonetization.

By using a standard transcription, important differences were observed depending on the corpus type (from 8.0% errors to 17.3% error rates using the rule-based system). The latter enrichment supposed that the pronunciation is standard, except if the manual transcription mentioned something else: the transcriber disambiguated pronunciations. Thus, this enrichment was particularly adapted to the automatic rule-based system. Indeed, the manual enrichment allowed the rule-based system to obtain a quite good phonetization. Furthermore, this paper proposed an algorithm that improved this phonetization by using a tree-based approach. The final system obtained a phonetization of about 95.2% correct (from 3.7% to 5.6% error rates depending on the corpus). Finally, the phonetization of Conversational Speech is as good as other types of corpora. Although if the enrichment is more time consuming, it constitutes therefore an effective alternative to phonetize this type of corpus. Such a transcription enrichment is necessary due to the fact that conversational data are still largely unknown.

7. Acknowledgements

This work was supported by ANR OTIM project Ref. Nr. ANR-08-BLAN-0239.

8. References

F. Bechet. 2001. LIA_PHON - un système complet de phonétisation de textes. *Traitement Automatique des Langues*, 42(1/2001).

R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. 2008. Le CID - Corpus of Interactional Data. *Traitement Automatique des Langues*, 49(3):105–134.

B. Bigi and D. Hirst. 2012. SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Speech Prosody*, Shanghai (China).

B. Bigi, C. Portes, A. Steuckardt, and M. Tellier. 2011. Multimodal annotations and categorization for political debates. In *ICMI Workshop on Multimodal Corpora for Machine learning*, Alicante (Spain).

P. Blache, R. Bertrand, and G. Ferré. 2009. *Creating and exploiting multimodal annotated corpora: the ToMA project*, volume LNAI 5509. Kipp M. (eds.).

P. Blache, 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, and S. Rauzy. 2010. Multimodal annotation of conversational data. In *The Fourth Linguistic Annotation Workshop*, Uppsala (Sueden).

P. Boersma and D. Weenink. 2009. Praat: doing phonetics by computer, http://www.praat.org.

ESTER2. 2008. Transcription détaillée et enrichie. convention d'annotation. http://www.afcp-parole.org/camp_eval_systemes_transcription/.

S. Herment, A. Loukina, A. Tortel, D. Hirst, and B. Bigi. 2012. A multi-layered learners corpus: automatic annotation. In *4th International conference on corpus linguistics Language, corpora and applications: diversity and change*, Jaén (Spain).

C. Meunier and R. Espesser. 2011. Vowel reduction in conversational speech in french: The role of lexical factors. *Journal of Phonetics. doi:10.1016/j.wocn.2010.11.008.*

Sclite. 2009. Speech Recognition Scoring Toolkit, http://www.itl.nist.gov/iad/mig/tools/, version 2.4.0.

S.J. Young and Sj Young. 1994. The HTK Hidden Markov Model Toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44.

AixOx - ex1	*I opened the front door to let the cat out*
AixOx - ex2	*send an ambulance to sixteen chadwick close as soon as possible*
Grenelle - ex1	*to replenish the bee population annually*
Grenelle - ex2	*beekeepers and particularly we do not know very well what is the cause of bee mortality*
	but there are still perhaps systemic attacks
CID - ex1	*thus he took the recipe and all well he said okay*
CID - ex2	*oh but that's just it, it was to sell you blablabla the guy pissed him off*
	and then he bought him the whatsit and then the guy left so i said shit, the guy wanted to...

Table 2: Orthographic Transcription: English translation

AixOx - ex1	j'ai ouvert la porte d'entrée pour laisser sortir le chat
AixOx - ex2	envoyer d'urgence une ambulance devant le numéro
	seize de l'impasse Claire Voie
Grenelle - ex1	à reconstituer leur cheptel d'abeilles tous les ans
Grenelle - ex2	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
CID - ex1	donc il prend la recette et tout bon il dit bon okay
CID - ex2	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

Table 3: Standard Orthographic Transcription

AixOx - ex1	j'ai ouvert la porte d'entrée pour laisser **chort-** sortir le chat
AixOx - ex2	envoyer d'urgence une ambulance devant le numéro
	seize de l'impasse Claire Voie
Grenelle - ex1	à reconstituer + leur cheptel d'abeilles tous les ans
Grenelle - ex2	**euh** les apiculteurs + et notamment **b-** on ne sait pas très bien
	+ quelle est la cause de mortalité des abeilles mais enfin y a
	quand même peut-être des attaques systémiques
CID - ex1	donc + **i-** il prend la **è-** recette et tout bon **il vé-** il dit bon okay
CID - ex2	ah mais justement c'était pour vous vendre bla bla bla **bl-** 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

Table 4: Orthographic Transcription with the first enrichment

AixOx - ex1	j'ai ouvert la porte d'entrée pour laisser **chort-** sortir le chat
AixOx - ex2	envoyer d'urgence une **[ambulance,ambulanceu]** devant **[le,leu]** numéro
	[seize,seizeu] de l' **[impasse,impasseu] [Claire Voie,claireuvoi]**
Grenelle - ex1	à **[reconstituer,reuconstituer]** + leur cheptel d'abeilles tous les ans
Grenelle - ex2	**euh** les apiculteurs + et notamment **b-** on **n(e)** sait pas très bien
	+ quelle est la cause de mortalité des abeilles **m(ais)** enfin y a
	quand même **peut-êt(r)e** des attaques systémiques
CID - ex1	donc + **i- i(l)** prend la **è-** recette et tout bon **i(l) vé- i(l)** dit bon **[okay, k]**
CID - ex2	ah mais justement c'était pour vous vendre bla bla bla **bl-** 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 **put(ain)** le mec **i(l)** voulait

Table 5: Orthographic Transcription with the second enrichment

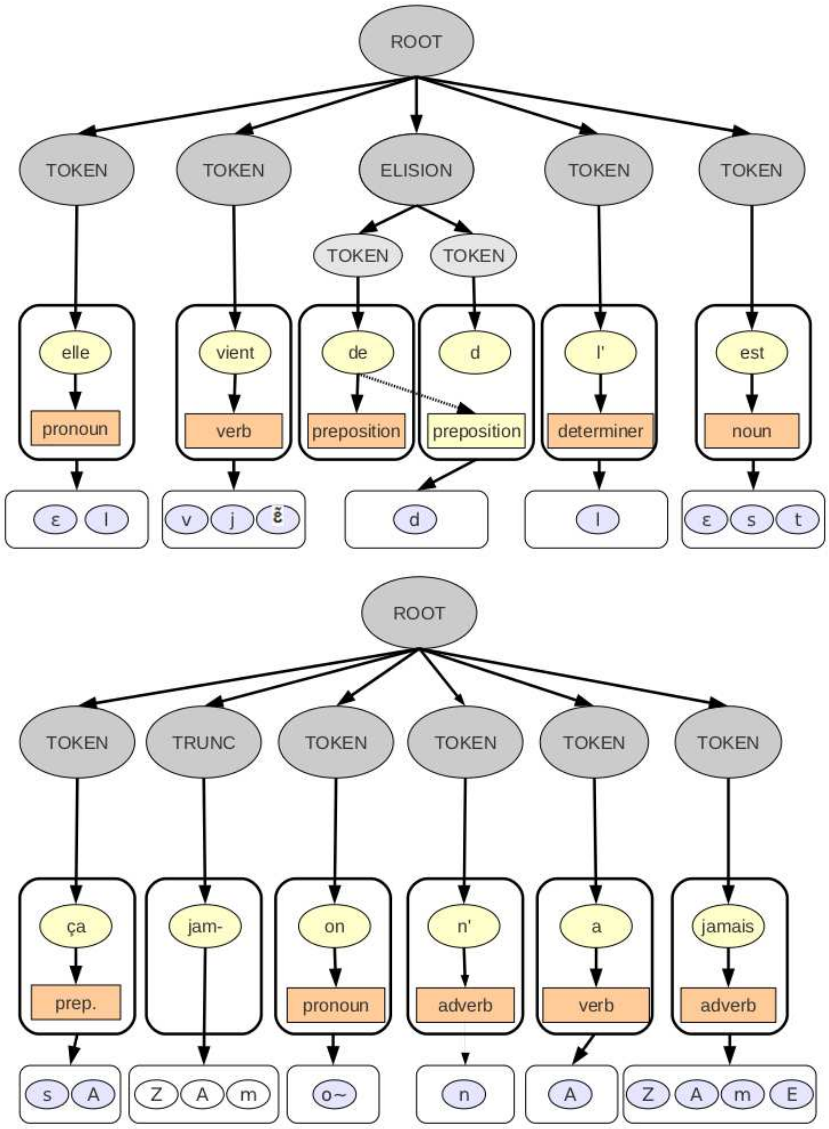

Figure 1: Tree-based phonetization examples