Influence de la transcription sur la phonétisation automatique de corpus oraux

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, peripauline@gmail.com, roxane.bertrand@lpl-aix.fr

RÉSUMÉ

Notre objectif vise à estimer l'influence de différents niveaux d'enrichissement de la transcription sur l'étape de phonétisation de l'oral. Cette étude a été réalisée sur un corpus test de 7 minutes, réparties entre trois types de données différentes (parole conversationnelle spontanée, lecture et discours politique). Les résultats montrent que plus la transcription bénéficie d'enrichissements, meilleure est la phonétisation obtenue, quel que soit le type de corpus.

ABSTRACT

what is the impact of the transcription on the phonetization

This paper aims at quantifying the impact of the transcription enrichments on the automatic phonetization of speech. Experiments were carried out on a 7 minutes French corpus including conversational speech, readed speech and a political discourse. Results showed the better the transcription the better the phonetization and that independently on the corpus.

MOTS-CLÉS : transcription, oral, enrichissement, phonétisation.

KEYWORDS: transription, phonetization, enrichment, speech.

1 Introduction

Pendant de nombreuses années, les transcriptions de corpus oraux étaient établies selon des conventions pouvant varier d'un auteur à l'autre, ou d'un projet à l'autre. Depuis une dizaine d'années, on constate de nombreux efforts de mutualisation et de partage des corpus. Ceci implique d'une part le recensement des différentes conventions existantes, d'autre part une tentative d'homogénéisation de ces conventions, quels que soient les objectifs et projets. Disposer de conventions communes permet de fournir aux transcripteurs des consignes précises qui contribuent surtout à rendre leurs transcriptions non seulement plus homogènes mais comparables et exploitables par une plus grande communauté d'utilisateurs. Le choix de certaines conventions mais plus encore celui des phénomènes à transcrire peut faciliter les traitements automatiques des corpus parmi lesquels les étapes de phonétisation, d'analyse morpho-syntaxique ou encore la reconnaissance automatique de la parole.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 449–456,
Grenoble, 4 au 8 juin 2012.

Dans cet article, nous abordons la question de la phonétisation des corpus oraux qui dépend de la transcription effectuée en amont. La phonétisation est l'étape consistant à convertir la suite de mots orthographiques en chaîne phonétique (ou en symboles phonétiques). Notre objectif est de mesurer l'influence des choix effectués lors de la transcription sur la phonétisation automatique.

Cet article se décline en 5 sections. La section 2 concerne la transcription du français oral et synthétise quelques conventions. La section 3 porte sur l'outil de phonétisation automatique utilisé. La section 4 présente le corpus de test. Enfin la section 5 expose la méthode et le résultat de l'évaluation.

2 Transcription

Il existe de nombreuses conventions de transcription en fonction des projets et objectifs de recherche. Nous ne visons pas l'exhaustivité mais nous avons sélectionné certaines conventions établies dans des projets relativements différents. Celles établies dans le cadre de la campagne ESTER visent à évaluer les systèmes de reconnaissance automatique de la parole (ESTER, 2008), tandis que celles du groupe ICOR (Groupe ICOR, 2007) portent essentiellement sur les interactions conversationnelles. Nous avons également examiné celles établies à l'ATILF (André *et al.*, 2009) et au centre de recherche Valibel (Bachy *et al.*, 2007). S'ajoute à cette liste les conventions établies au LPL dans le cadre du projet ANR OTIM (Blache *et al.*, 2010), inspirées des conventions du GARS (Blanche-Benveniste et Jean-Jean, 1987).

Le tableau 3 synthétise les notations des différentes conventions pour les phénomènes propres à l'oral. Les cases vides signifient que le phénomène n'est pas mentionné dans la convention de transcription. On remarque que certaines conventions, comme celle d'ICOR, sont plus proches de ce qui a été prononcé (en particulier pour les élisions) alors que d'autres conventions (en particulier Valibel) s'orientent vers une orthographe standard. La convention du LPL propose une double orientation : en mentionnant les élisions entre parenthèses par exemple, un traitement automatique peut retrouver soit l'orthographe standard (en supprimant seulement les '()') soit la prononciation (en supprimant les '()' et leur contenu).

Notre étude porte exclusivement sur les aspects des transcriptions qui sont susceptibles d'affecter la phonétisation. Le but de cet article est d'estimer l'influence des enrichissements de la transcription orthographique sur la phonétisation automatique de l'oral, quels que soient les symboles utilisés pour les transcrire.

L'examen des similitudes et des différences de ces conventions, a servi de support pour définir 3 niveaux d'enrichissements sur lesquels portera l'évaluation (cf tableau 3).

3 Phonétisation

En dehors de l'étape de transcription graphème-phonème, généralement traitée par une approche à base de règles, de nombreux traitements linguistiques sont nécessaires afin de lever les ambiguïtés de prononciations. Parmi celles-ci, citons les problèmes liés au formatage du texte, aux homographes hétérophones, aux liaisons, à la phonétisation des noms propres, des sigles ou des emprunts à des langues étrangères. L'outil LIA_Phon (Bechet, 2001), qui a été utilisé dans

la présente étude, considère l'ensemble de ces cas. Le choix de cet outil est, en outre, lié à ses conditions d'accessibilité, sa bonne documentation et surtout ses performances.

Les outils inclus dans le LIA_Phon peuvent se décomposer en trois modules : les outils de formatage et d'étiquetage (LIA_Tagg), les outils de phonétisation et les outils d'exploitation des textes phonétisés. Dans la présente étude, nous faisons appel aux deux premiers modules. Les outils de formatage et d'étiquetage permettent de traiter le texte brut à phonétiser. Cet ensemble d'outils regroupe des modules de découpage (en mots et en phrases), de correction (traitement des capitalisations, des formes désaccentuées et des abréviations) et d'étiquetage (morphologique et syntaxique). À la suite de ces traitements, la plupart des ambiguïtés de prononciation sont levées. Le module de phonétisation regroupe d'une part un ensemble de bases de règles de phonétisation relatives aux étiquettes préalablement posées et d'autre part un module de traitement des liaisons gérant les liaisons interdites, facultatives et obligatoires.

Un module a été spécifiquement développé pour transformer la transcription enrichie en une chaîne de caractères prête à être utilisée par le LIA_Phon.

4 Corpus de test

À notre connaissance, il n'existe pas de corpus phonétisé manuellement qui soit disponible publiquement afin d'évaluer les phonétisations automatiques. Nous avons donc construit un tel corpus que nous avons déposé sur la forge Speech Language Data Repository (SLDR) [1]. Ce corpus [2] dure environ 7 minutes. Les durées et autres détails sont décrits dans le tableau 1. Il contient des extraits des corpus suivants :

– CID [3], corpus conversationnel décrit dans (Bertrand *et al.*, 2008),
– AixOx [4], corpus de lecture décrit dans (Herment *et al.*, 2012),
– Grenelle [5], intervention d'Yves Cochet lors d'un débat à l'Assemblée nationale portant sur le « Grenelle II de l'environnement » décrit dans (Bigi *et al.*, 2011).

Le corpus MARC-Fr a été entièrement phonétisé et aligné manuellement par un expert phonéticien. Pour illustrer les phénomènes reportés dans le tableau 1, quelques exemples sont reportés ci-après en respectant les conventions d'écriture des transcriptions du LPL, à savoir :

– les amorces sont notées avec un tiret collé à la fin,
– les pauses perçues et inférieures à 200 ms sont notées "+",
– les élisions non standards mentionnent entre parenthèse ce qui n'est pas prononcé,
– les prononciations non standards sont spécifiées entre crochets, avec en partie gauche l'orthographe standard et en partie droite la réalisation effective.

Deux exemples du CID sont présentés ci-après. Comme on le voit dans le tableau 1, en tant que corpus conversationnel, celui-ci comporte de nombreux phénomènes tels que les amorces, ou les pauses pleines. De plus il contient de nombreux phénomènes de réduction, notamment des déformations, des assimilations, des élisions de phonèmes, qui s'avèrent extrêmement fréquents en parole naturelle non contrôlée.

1. http ://www.sldr.fr/
2. MARC-Fr : dépôt SLDR numéro 000786
3. CID : dépôt SLDR numéro 000027
4. AixOx : dépôt SLDR numéro 000784
5. Grenelle : dépôt SLDR numéro 000729

	CID	AixOx	GrenelleII
Durée	143s	137s	134s
Nombre de locuteurs	12	4	1
Nombre de phonèmes	1876	1744	1781
Nombre de mots	1269	1059	550
Pauses perçues	10	23	28
Pauses pleines	21	0	5
Bruits (souffles,...)	0	8	0
Rires	4	0	0
Amorces	6	2	1
Élisions non standards	60	21	34
Prononciations particulières	58	37	23

TABLE 1 – Description du corpus de test MARC-Fr

```
1/ donc + i- i(l) prend la è- recette et tout bon i(l) vê- i(l) dit bon
[okay, k]
2/ ouais tu comprends na na na na na la solidarité les étudiants
et [quelle, qué] solidarité ah c'est bon j(e) [lui,i] dis [tu, ty] es
solidaire toi t'es [solidaire,solidaireu] [de,deu] [de,deu] [de,deu] tes
[fesses,fèsseu] t'es solidaire
```

Voici ensuite des exemples du corpus AixOx. Ce corpus lu comprend un très petit nombre d'amorces, d'hésitations et quelques élisions non standard. Néanmoins, il contient un nombre assez important de prononciations particulières, qui proviennent de l'accent très marqué de l'un des locuteurs (exemple 3).

```
1/ j'ai ouvert la porte d'entrée pour laisser chort- sortir le chat
2/ l'un [des,nèn] deux l'un des deux individus en état d' ébriété a été
appréhendé
3/ envoyer d' urgence une [ambulance,ambulanceu] devant [le,leu] numéro
[seize,seizeu] de l' [impasse,impasseu] [Claire Voie,claireuvoi]
```

Enfin, deux exemples du corpus Grenelle sont reportés. Il est intéressant de noter qu'au début du second exemple, Yves Cochet est interrompu par des remarques des députés, ce qui explique les pauses et hésitations.

```
1/ à [reconstituer,reuconstituer] + leur cheptel d'abeilles tous les ans
2/ 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 euh peut-êt(r)e des attaques systémiques
```

La transcription du corpus de test a été effectuée avec le logiciel Praat (Boersma et Weenink, 2009), selon les conventions du LPL. Bien que le temps de transcription soit variable d'un corpus à l'autre, d'un annotateur à l'autre, ou encore d'un outil à l'autre, tenter d'estimer le temps/coût d'une transcription s'avère particulièrement utile. La transcription s'est déroulée en 3 étapes. La première étape a consisté à transcrire orthographiquement, en ajoutant les pauses silencieuses,

pauses perçues, rires et amorces. Le temps de cette transcription a varié entre 12 et 20 minutes par minute de parole selon le corpus considéré (plus de temps pour le CID, moins pour le Grenelle). La deuxième étape consistait à ré-écouter et ajouter les élisions et prononciations particulières. Pour cette étape, le temps a varié entre 10 et 20 minutes par minute de parole et ce davantage en raison du locuteur que du corpus lui-même : les locuteurs ayant un accent régional fortement marqué ont nécessité plus de temps que les autres. Enfin, la troisième écoute a consisté simplement à vérifier la version produite, et a nécessité en moyenne 10 minutes par minute de parole (temps relativement constant sur le corpus). Dans tous les cas, au moins deux personnes sont intervenues sur la transcription (en réalisant l'une des trois étapes). Il est important en effet que la transcription fasse l'objet d'au moins une vérification systématique par une autre personne que le transcripteur initial.

5 Évaluations

Les évaluations ont été effectuées avec l'outil Sclite (Speech Recognition Scoring Toolkit, 2009). Habituellement utilisé en reconnaissance automatique de la parole où il estime un Taux d'Erreurs Mots, il calcule ici un Taux d'Erreurs Phonèmes (Err) qui somme les erreurs de :
– substitution (Sub), exemple : UN / AI
– suppression (Del), exemple : pp EU tt ii / pp tt ii
– insertion (Ins), exemple : jj / jj EU
Pour les évaluations, nous avons utilisé un jeu de phonèmes réduit, en combinant les paires suivantes : oo/au, ei/ai, yy/ii. Ces 3 fusions concernent environ 2,7% (absolu) des erreurs par substitution, quels que soient le corpus et la transcription. Les liaisons ne sont pas "traitées" ici : dans tous les cas, on utilise les liaisons obligatoires proposées par le LIA_Phon.

Nous avons comparé trois types de transcription :

1. la transcription orthographique (TO) standard ;
2. la TO enrichie - 1 qui contient un dénominateur commun aux enrichissements proposés par les différentes conventions, à savoir les pauses perçues, les pauses pleines, les répétitions disfluentes, les rires, les bruits, les amorces (équivaut aux enrichissements de la convention ESTER) ;
3. la TO enrichie - 2 qui ajoute à la précédente les élisions non standards (présentes dans les conventions ICOR et LPL) et les prononciations dites particulières (présentes dans les conventions TCOF, VALIBEL et LPL).

Les résultats sont présentés dans le tableau 2. On observe que la phonétisation obtenue à partir de l'orthographe standard est très éloignée de celle attendue, quel que soit le corpus, mais de façon significative pour les données du CID. L'enrichissement (phonétique) 1, apporté par l'ensemble des conventions, permet un gain important : 3,2 % pour les corpus CID et AixOx mais seulement 1,7 % pour le corpus Grenelle où Yves Cochet intervient à l'Assemblée nationale. L'enrichissement 2, qui mentionne les élisions non standards et les prononciations dites particulières, permet de produire une phonétisation significativement meilleure pour chacun des 3 corpus. Il divise même par deux le nombre d'erreurs pour le CID.

L'analyse de détail des erreurs révèle un grand nombre d'insertions pour la transcription orthographique standard, en particulier pour le CID qui contient un grand nombre de phénomènes liés à la réductions de la parole. On observe aussi beaucoup de suppressions (Del) car il manque

	Sub %	Del %	Ins %	Err %
CID				
TO standard	2,8	4,5	10,0	17,3
TO enrichie - 1	2,7	1,4	10,3	14,4
TO enrichie - 2	1,8	1,3	3,4	6,5
AixOx				
TO standard	1,4	5,0	3,0	9,5
TO enrichie - 1	1,4	2,3	2,9	6,5
TO enrichie - 2	1,3	1,8	2,5	5,6
Grenelle				
TO standard	1,1	2,8	4,1	8.0
TO enrichie - 1	1,0	1,2	4,1	6,3
TO enrichie - 2	1,3	1,0	1,7	4,0

TABLE 2 – Pourcentages d'erreurs de la phonétisation

à cette transcription tous les phénomènes propres à l'oral qui n'ont donc pas été phonétisés. La transcription enrichie 1 permet ainsi de diminuer significativement le nombre de suppressions. Il reste toutefois beaucoup de suppressions dans le corpus AixOx ; elles sont dues à l'accent d'un des locuteurs qui prononce les schwas finaux et ne produit pas les élisions standards. Cet enrichissement n'a cependant pas d'impact sur les erreurs d'insertions ou les substitutions par rapport à une TO standard. L'enrichissement 2 permet de réduire significativement le taux d'erreurs d'insertions, en particulier pour le CID où il est divisé par 3 et pour le Grenelle où il est divisé par 2,5. La transcription enrichie 2 permet aussi de réduire le taux de suppressions dans le cas du corpus AixOx.

6 Conclusion

Cet article a évalué l'influence que le niveau d'enrichissement des transcriptions peut avoir sur la qualité de la phonétisation automatique de corpus oraux. Les résultats confirment que les enrichissements contribuent à améliorer la phonétisation et ce quel que soit le type de corpus : lecture, discours, conversationnel. L'amélioration est bien entendu particulièrement significative pour ce dernier qui présente davantage de phénomènes non standards (parole non préparée). Bien que plus coûteux en temps, l'enrichissement manuel permettant d'obtenir une phonétisation de qualité quasiment égale à celle obtenue pour les autres corpus, constitue donc une alternative efficace pour phonétiser ce type de corpus. Une telle transcription (très riche) s'est avérée nécessaire en raison du fait que les données conversationnelles étaient encore largement méconnues. Mais avec le partage des corpus, la volonté d'établir des conventions communes et des études comparatives telles que celles présentées ici, on peut envisager à terme de mieux recenser et décrire les phénomènes propres aux différents corpus en vue de les intégrer directement via des étapes plus automatiques.

Références

André, V., Benzitoun, C., Canut, E., Debaisieux, J.-M., Gaiffe, B. et Jacquey, E. (2009). Conventions de transcription en vue d'un alignement texte-son avec transcriber. TCOF : Traitement de corpus oraux en français, ATILF, Nancy, http ://www.cnrtl.fr/corpus/tcof/.

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.-C. et Wilmet, R. (Version revue en juin 2004 ; mise à jour : 18/04/2007). Conventions de transcription régissant les corpus de la banque de données valibel. Université catholique de Louvain, http ://www.uclouvain.be/81836.html.

Bechet, F. (2001). Lia_phon - un systeme complet de phonetisation de textes. *Traitement Automatique des Langues*, 42(1/2001).

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

Bigi, B., Portes, C., Steuckardt, A. et Tellier, M. (2011). Catégoriser les réponses aux interruptions dans les débats politiques. *In 18èmes conférence annuelle Traitement Automatique des Langues Naturelles*, pages 167–172, Montpellier (France).

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

Blanche-Benveniste, C. et Jean-Jean, C. (1987). *Le français parlé. Transcription et édition*. Paris : Didier érudition.

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

ESTER (version 0.1 - 08/01/2008). Ester2 : Transcription détaillée et enrichie. convention d'annotation. http ://www.afcp-parole.org/camp_eval_systemes_transcription/.

Groupe ICOR, I. C. L. . E.-L. (Mise à jour : novembre 2007). Convention icor. http ://clapi.univ-lyon2.fr/.

Herment, S., Loukina, A., Tortel, A., Hirst, D. et Bigi, B. (2012). A multi-layered learners corpus : automatic annotation. *In 4th INTERNATIONAL CONFERENCE ON CORPUS LINGUISTICS Language, corpora and applications : diversity and change*, Jaén (Espagne).

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

	ESTER	ICOR	TCOF	VALIBEL
Incompréhen- -sible	[pron=pi]	autant de 'x' que de syllabes	autant de '*' que de syllabes	
Inaudible	[pron=pi]	(inaud.)	'*'	
Élisions	orthographe *il y a déjà*	graphie substituée par ' *i' y a d'jà*	orthographe *il y a déjà*	orthographe *il y a déjà*
Troncations, Amorces	insertion de () *car()*	insertion d'un '-' *car-*	insertion d'un '-' *car-*	insertion d'un '/' *car/*
Amorces avec continuation	orthographe std *c'est incroyable*	*c'est in- croyable*		*c'est in/ croyable*
Prononciations particulières	orthographe std commence par '*' *qu'il *soit là*		entre [] après la graphie *qu'il soit [pron=swaj] là*	entre [] après la graphie *qu'il soit [swaj] là*
Liaisons particulières			graphie entre '=' *le =n= ours*	phonème entre '[]' *donne moi [z] en*
Pauses	()	(.) si < à 0,2 s (durée) sinon	'+' très longues '///'	brève '/' longue '//'
Rires	[b]		[rire]	(rire)
Toux	[b]			(toux)
Soupir	[b]			(soupir)
Bâillement				(baillement)
Inspiration	[r]	.h :		(inspiration)
Expirations	[r]	h : :		(expiration)
Bruit	[b]		[bruit]	(bruit)

TABLE 3 – Conventions de transcription de phénomènes de l'oral