# Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage

*M. Adda-Decker, B. Habert, C. Barras, G. Adda, Ph. Boula de Mareüil & P. Paroubek*

LIMSI-CNRS
BP 133 – 91403 Orsay Cedex, France
Tél. : ++33 (0)1 69 85 80 06 – Fax : ++33 (0)1 69 85 80 88
Mél : {madda,habert,barras,gadda,mareuil,pap}@limsi.fr - http://www.limsi.fr/

## ABSTRACT

The aim of this study is to elaborate a disfluent speech model by comparing different types of audio transcripts. The study makes use of 10 hours of French radio interview archives, involving journalists and personalities from political or civil society. A first type of transcripts is press-oriented: most disfluencies are discarded. For 10% of the corpus, we produced exact audio transcripts: all audible phenomena and overlapping speech segments are transcribed manually. In these transcripts, about 14% of the words correspond to disfluencies and discourse markers. The audio corpus has then been transcribed using the LIMSI speech recogniser. With 8% of the corpus, the disfluency words explain 12% of the overall error rate. This shows that disfluencies have no major effect on neighbouring speech segments. Restarts are the most error prone, with a 36.9% within class error rate.

## 1. INTRODUCTION

Au sein du département Communication Homme-Machine du LIMSI existent des compétences en traitement automatique de l'écrit comme de l'oral que nous cherchons à combiner. Les documents textuels et audio traitant d'un même thème sont des ressources de plus en plus accessibles, qui peuvent dessiner un premier axe de recherche. Ces documents peuvent simplement aborder un domaine commun (informations télévisées et articles de presse, par exemple) ; les textes peuvent aussi, comme dans la présente étude, être des transcriptions relativement fidèles de données de parole. Ils permettent alors d'adapter les modèles de langage et les lexiques spécifiques. Une meilleure connaissance des phénomènes de parole spontanée est également nécessaire pour améliorer la modélisation : hésitations, lapsus, faux départs, répétitions (de mots-outils), etc., que nous appellerons « disfluences ». Un deuxième axe de recherche consiste à enrichir la transcription de parole pour la donner en entrée à des étiqueteurs/analyseurs syntaxiques et à des outils d'indexation. Comme ces logiciels peuvent avoir besoin de signes de ponctuation pour faciliter la détection de phrases, de propositions et de syntagmes, des modèles linguistiques de la ponctuation et de frontières acoustiques telles que l'allongement final, les inspirations, les disfluences et l'intonation peuvent être utilisés.

Nous nous appuyons ici sur 10 heures d'archives télévisées françaises enregistrées il y a environ 10 ans : 10 émissions longues d'une heure où une personnalité politique ou représentant la société civile (par exemple une ONG) est interrogée par plusieurs journalistes. Ces derniers ont préparé leurs questions, on pouvait en deviner la plupart, et les réponses ne sont pas entièrement spontanées. Mais d'autre part, l'un des journalistes dirige les débats, surveille le temps consacré aux différents thèmes (déterminés au préalable), et interrompt souvent la personne interviewée ou posant les questions. Cette configuration favorise les disfluences, et les chevauchements de parole sont fréquents. Seulement une partie des nombreuses disfluences fournit de l'information quant à l'idéation du locuteur [5] ; le reste revient à une lutte pour la parole entre interlocuteurs, même si les journalistes ne « coupent » pas la parole n'importe quand [6].

Pour chaque émission, nous disposons à la fois de l'audio et d'une transcription destinée à la presse (TPress). Ces transcriptions se veulent assez proches de l'audio (des citations en sont extraites dans d'autres média) tout en suivant des conventions implicites : elles se situent quelque part entre texte écrit et transcription exacte. De fait, la plupart des disfluences et autres fautes de langue ont été ignorées. Aussi avons-nous produit manuellement une transcription exacte (TExact) pour 10 % des données audio, avec tous les phénomènes audibles, notamment en parole superposée Pour ce faire, nous nous sommes aidés des transcriptions automatiques (TReco), car il est facile de « manquer » inconsciemment certaines disfluences.

La comparaison des transcriptions TPress, TExact et TReco soulève les questions suivantes : quelle est la proportion globale de disfluences observées ? comment se répartissent les différents types de disfluences ? cette répartition est-elle corrélée avec les caractéristiques sociologiques des locuteurs ou avec leur compétition pour la parole de « premier plan » ? Est-ce que certaines classes de disfluences sont mieux détectées que d'autres ? Est-il difficile d'en rendre compte à l'aide de *n*-grammes de mots conventionnels ?

## 2. ANNOTATION DE LA PAROLE SPONTANÉE

La parole spontanée a suscité l'intérêt de plusieurs

équipes francophones. Morel et Danon-Boileau [4] ont abordé ces « "petits mots" qui habitent l'oral », qu'ils nomment « ligateurs » : ex. *quoi, bon, enfin, donc, alors, genre, style*. Le GARS, à Aix-en-Provence [2], a travaillé pendant des années sur les problèmes que pose la représentation écrite de l'oral, avec en vue une exploitation grammaticale : les choix sont un compromis entre fidélité et lisibilité, d'où une transcription en orthographe standard. Un autre projet, PFC (Phonologie du Français Contemporain [3]), visant à couvrir un vaste espace géographique, a pour objectif d'aligner aussi facilement que possible les données de parole avec des textes écrits.

Notre travail suit plutôt les directives d'annotation metadata du LDC [7], adoptées dans les évaluations Rich Transcription (RT) conduites par le NIST (http://www.nist.gov/speech/tests/rt/rt2003/). Ayant de même pour but de produire des transcriptions aussi lisibles que possible, nous avons choisi ces conventions parce qu'elles sont compatibles avec nos propres objectifs et qu'elles représentent le résultat d'une vaste discussion. Les annotateurs doivent indiquer ce qui fonctionne pragmatiquement comme une unité de pensée cohérente (noté SU). Ils distingueront ensuite les remplisseurs comme *euh* (FW : *Filler Word*), ce qui structure le discours (DM : *Discourse Marker*), la reconnaissance explicite d'une disfluence par le locuteur (EET : *Explicit Editing Term*), les apartés (AS : *ASide*) et les incises portant, contrairement aux remarques ou commentaires précédents, sur le même thème que l'énoncé principal (PA : *PArenthetical*). D'autres disfluences seront notées RP (*RePetition*), RV (*ReVision*) ou RS (*ReStart*, quand une correction, remplaçant ce qui précède, en modifie le sens).

Pour annoter notre corpus d'archives, nous avons décidé d'adapter les directives du LDC au français avec quelques simplifications. Nous marquons les PA et AS dans la transcription exacte, sans les commenter. Et nous avons rassemblé RV et RS sous le chapeau RS, car il n'est pas toujours facile d'évaluer la modification de sens voulue.

## 3. CORPUS ET TRANSCRIPTION

### *3.1. Corpus et transcription exacte*

Par la suite, il est fait référence à chaque locuteur par un chiffre de 1 à 20 suivi de lettres fournissant quelques unes de ses caractéristiques sociologiques, comme le montre la table 1. Si rien de plus n'est précisé, un locuteur est un Français — il n'y a qu'une femme parmi les interviewés, qui sont par défaut des hommes politiques. L'un d'entre eux est anglais ; deux autres sont des francophones issus de pays africains.

Une transcription exacte de l'audio a été produite manuellement sur 10 % du corpus (soit 1 heure de parole) : 2 extraits d'environ 3 minutes, sélectionnés aléatoirement dans chaque émission, ont été divisés en SU, et toutes les disfluences ont été explicitées et annotées comme détaillé précédemment. Le nombre de mots par SU va de 8,6 à 20,8 (médiane : 12,8 ; moyenne : 13,7). Les médiane et moyenne sont plus élevées chez les interviewés que chez les journalistes (médiane : 13,9/11,6 ; moyenne : 14,9/12,2). Pour caractériser les locuteurs selon leurs disfluences, une analyse de correspondance a été utilisée pour les traits DM, RP, RS et FW ; mais elle ne montre aucun regroupement soit des journalistes soit des interviewés.

**Table 1 :** Sens des codes et nombres de locuteurs.

| code | sens | #loc. | code | sens | #loc. |
|---|---|---|---|---|---|
| **J** | journaliste | 9 | **I** | interviewé | 11 |
| **C** | chairman | 1 | **w** | femme | 1 |
| **e** | natif anglais | 1 | **o** | personne âgée | 1 |
| **r** | *accent* région. | 1 | **c** | société civile | 2 |
| **f** | francophone | 2 | | | |

Comme la figure 1 le montre, pour 13 locuteurs (7 journalistes et 6 interviewés), la proportion des RP est supérieure à celle des RS au sein des disfluences ; mais 5 interviewés (11-Ic, 15-I, 16-Io, 18-If, 20-Ie) et un journaliste (6-J) montrent le contraire. L'utilisation des RP est clairement dominante chez les journalistes, peut-être en raison des difficultés que ceux-ci rencontrent quand ils essaient d'interrompre la personne qu'ils interrogent. En revanche, les RS l'emportent pour la moitié des interviewés, qui semblent avoir de réelles occasions de chercher leurs mots.

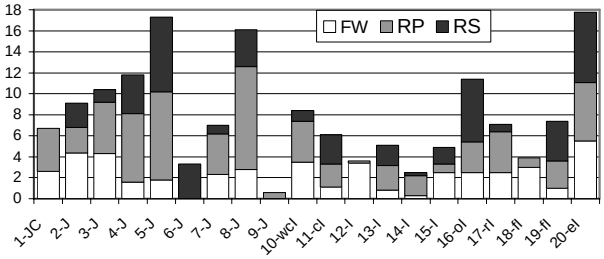


**Figure 1 :** Proportions des disfluences (FW, RP et RS) dans les transcriptions TExact de chaque locuteur.

### *3.2. Transcription automatique de l'audio*

Le corpus audio a été transcrit en utilisant un système de reconnaissance de la parole, aboutissant aux transcriptions TReco. Le système standard du LIMSI pour les informations télé ou radiodiffusées en français [1] a été utilisé pour transcrire une heure du corpus. Les modèles acoustiques ont été entraînés sur approximativement 100 heures de parole provenant d'informations télé ou radiodiffusées en français : ils consistent en des modèles dépendants du contexte de 33 phonèmes, plus 3 modèles génériques pour silences, hésitations et respirations. Le modèle de langage standard est une interpolation de modèles 4-grammes entraînés sur différents ensembles de données. Trois

sources ont été utilisées : transcriptions destinées à la presse de diverses émissions de radio ou de télévision (48 millions de mots), des transcriptions exactes d'émissions de radio principalement (950 000 mots) et des articles de journaux (311 millions de mots). Le lexique contient 65 000 mots, choisis pour optimiser la couverture des données de développement (télé/radio très différentes en date et en source du corpus d'archives de cette étude). Les prononciations sont dérivées d'un convertisseur graphème-phonème par règles, et vérifiées manuellement. Le système tourne à environ 10 fois le temps réel sur un PC standard.

En utilisant les transcriptions pour la presse fournies avec le corpus (580 000 mots), un modèle de langage « informé » a été construit par interpolation avec le modèle *n*-gramme standard. Le lexique contient seulement les 26 000 mots les plus fréquents dans les sources standard, ainsi que les 19 000 mots contenus dans les transcriptions destinées à la presse, aboutissant à un lexique de 30 000 mots.

La performance de la transcription automatique de la parole a été évaluée en utilisant l'outil NIST `sclite`, en comptant le pourcentage de mots différents par rapport à la transcription TExact. Les disfluences ont été étiquetées dans la référence comme mots optionnels (i.e. aucune erreur n'était comptée si une pause sonore, une répétition ou un faux départ était ignoré par le système). La parole superposée, où les locuteurs parlent vraiment en même temps, a pour l'essentiel été écartée de l'évaluation ; cependant, il reste une quantité non négligeable de bruit ou de parole de fond, ce qui est la principale source d'erreurs de la part du système.

En utilisant le système de transcription standard pour le français, un taux d'erreurs moyen de 24 % sur les mots pouvait être mesuré. Ce pourcentage relativement haut doit être comparé avec deux autres chiffres : le taux d'erreurs standard sur d'autres données télé ou radiodiffusées en français (environ 20 % sur les mots) et le résultat actuel de la dernière évaluation RT sur des données comparables en anglais américain (11,7 %). À partir de cette comparaison, on peut s'attendre à des progrès en travaillant sur les spécificités du français. On peut aussi développer des modèles acoustiques dédiés à notre corpus d'archives, pour réduire le fossé avec notre taux d'erreurs standard en français.

Dans une seconde expérience, le modèle de langage informé a été utilisé : le taux d'erreurs résultant est de 14,5 % (1365 mots sur 9400), soit une réduction relative de 40 %, comparé avec les 24 % obtenus avec le système standard. Un des buts de cette expérience avec des transcriptions informées était de tester si des transcriptions précises peuvent être obtenues en partant de transcriptions pour la presse, rapides à produire. Le haut taux d'erreurs sur les mots montre qu'alimenter simplement le modèle de langage de transcriptions pour la presse ne suffit pas à produire des transcriptions de haute qualité. Il met également en évidence des erreurs qui viennent principalement de problèmes acoustiques. Les résultats par locuteur sont donnés dans la table 2, et révèlent une large variabilité inter-locuteur.

**Table 2 :** Taux d'erreurs sur les mots du système standard (S) et du système informé (I).

| Journaliste | S | I | Interviewé | S | I |
|---|---|---|---|---|---|
| 1-CJ | 33,0 | 22,7 | 10-Iwc | 24,2 | 14,2 |
| 2-J | 19,7 | 13,3 | 11-Ic | 25,5 | 13,6 |
| 3-J | 16,9 | 10,1 | 12-I | 17,4 | 4,9 |
| 4-J | 23,7 | 11,2 | 13-I | 19,8 | 10,3 |
| 5-J | 25,8 | 17,1 | 14-I | 16,6 | 8,6 |
| 6-J | 18,8 | 6,0 | 15-I | 16,7 | 9,8 |
| 7-J | 36,2 | 23,8 | 16-Io | 35,0 | 21,2 |
| 8-J | 24,6 | 23,8 | 17-Ir | 27,8 | 16,7 |
| 9-J | 14,0 | 3,0 | 18-If | 28,4 | 15,5 |
| **Tous** | **24,0** | **14,5** | 19-If | 32,7 | 24,4 |
| | | | 20-Ie | 28,7 | 22,5 |

## 4. COMPARAISON DES TRANSCRIPTIONS MANUELLES

Les transcriptions destinées à la presse sont assez proches des données audio. Pour se donner une idée des différences entre les versions TPress et TExact, `sclite` est à nouveau utilisé avec, comme référence, la version TExact où toutes les disfluences ont été filtrées. Le taux de mots différents monte à 9 %. Même si les disfluences ne sont évidemment pas la seule cause de différence entre les deux versions, nous allons dans ce qui suit nous concentrer sur les 3 types principaux de disfluence : FW, RP et RS. La classe FW contient un seul élément (*euh*). La majorité des répétitions (RP) est de la forme la plus simple : deux monosyllabes consécutifs. De bons candidats aux répétitions sont les déterminants, les pronoms, les prépositions et les adverbes, les items les plus observés étant *le, les, un, qui, que, de, à, très, pas, et*. Bien sûr, on observe des configurations de répétitions plus complexes (~~*beaucoup de*~~*, beaucoup de ;* ~~*peut-être*~~ *alors peut-être ;* ~~*et et et et et le plus*~~ *et le plus*…), mais cela ne représente qu'un faible pourcentage des répétitions. La classe RS est la plus hétérogène : des faux départs peuvent être simplement dus à l'anticipation d'un article erroné dans sa forme ou son genre, qui nécessite une correction (~~*pour le*~~ *pour l'événement*). Mais outre cette catégorie simple, tout syntagme peut être corrigé ou abandonné, et il est impossible d'en donner un aperçu synthétique.

Environ 8 % des mots de TExact sont de l'un de ces types FW (2,5 %), RP (3,2 %) ou RS (2,3 %). En outre, 6,3 % des mots correspondent à des marqueurs de discours (DM), qui ne sont pas réellement des disfluences, mais sont typiques du langage parlé. Leur rôle consiste plus ou moins à amorcer un tour de parole et à lier les énoncés entre eux. Ces connecteurs semblent particulièrement utiles en situation de lutte pour la parole, et impliquent généralement un nombre limité de mots : *alors, et, mais, donc, bon, voilà, oui, hein*. Cependant, chaque locuteur peut avoir ses propres préférences et habitudes en la matière.

## 5. ERREURS DE RECONNAISSANCE : LE RÔLE DES DISFLUENCES

Afin de mesurer la contribution des disfluences au taux d'erreurs global, la table 3 montre les sources d'erreurs majeures, en commençant par les classes présentées ci-dessus et celle des marqueurs de discours. Outre ces événements, la prononciation réduite de mots ou de séquences de mots courants est une source importante d'erreurs : tandis que les disfluences seules rendent compte d'environ 12,5 % des erreurs observées, les DM produisent 8,2 % d'erreurs. Une part plus importante de 25,1 % vient de prononciations réduites.

**Table 3 :** Nombre et pourcentage global d'erreurs observées dans différentes classes — la première correspondant aux disfluences, la dernière aux prononciations réduites (PR, parole rapide et mal articulée).

| *Classe* | *#erreurs* | *%erreurs global* |
|---|---|---|
| FW+RP+RS | 171 | 12,5 % |
| FW+RP+RS+DM | 283 | 20,7 % |
| PR | 347 | 25,1 % |

Il est également intéressant de savoir si les disfluences engendrent plus d'erreurs que d'autres mots. La table 4 montre le nombre d'erreurs, le taux d'erreurs correspondant et le nombre total de mots dans chaque classe. Tandis que tous les taux d'erreurs sont au-dessus du taux d'erreurs moyen du corpus, on voit que certaines classes sont plus difficiles à traiter que d'autres : 36,9 % d'erreurs pour RS contre 15,3 % pour RP. On observe également des différences significatives entre locuteurs. Parmi les interviewés, un locuteur non natif produit la moitié de toutes les erreurs sur les répétitions (23 erreurs). En l'excluant juste des comptes, le taux d'erreurs sur les répétitions tombe à 8,8 %, ce qui est bien inférieur au taux d'erreurs moyen (13,8 % sans ce locuteur non natif).

**Table 4 :** Taux d'erreurs pour différentes classes.

| *Classe* | *#err.//#total* | *%err. dans la classe* | *%err. global* |
|---|---|---|---|
| FW | 45 / 231 | 19,5 % | 3,0 % |
| RP | 46 / 300 | 15,3 % | 3,0 % |
| RS | 80 / 217 | 36,9 % | 6,5 % |
| DM | 112 / 593 | 19,3 % | 8,2 % |

## 6. DISCUSSION

Dans cette étude, nous avons comparé différents types de transcriptions audio avec, comme objectif, une meilleure modélisation des spécificités de la parole spontanée et de leur rendu. La comparaison de transcriptions destinées à la presse et exactes de l'audio a montré que les disfluences expliquent seulement la moitié des différences. La moitié restante revient principalement à la transcription des marqueurs de discours, des incises, de la parole retravaillée ou superposée. Tandis que de nombreuses disfluences peuvent simplement être filtrées dans les transcriptions, d'autres sont porteuses d'information : les hésitations peuvent indiquer des ruptures syntaxiques, et les garder accroît la lisibilité. D'autres types de disfluences sont plus délicats à supprimer, ce qui est également le cas de certains connecteurs pragmatiques qui doivent être remplacés par leur équivalent dans la langue écrite.

Concernant la transcription automatique, nous avons étudié l'impact des disfluences sur les taux d'erreurs sur les mots. Avec 8 % du corpus, ces disfluences expliquent 12 % du taux d'erreurs global, ce qui suggère qu'elles n'exercent pas un effet majeur sur les segments voisins. Les faux départs sont le plus objets à erreurs, avec un taux d'erreur de 36,9 %. Toutefois, les manipuler à un simple niveau lexical s'avère insuffisant : introduire des informations morpho-syntaxiques peut ici fournir un niveau de modélisation utile. Si l'on excepte la parole superposée, les prononciations réduites se révèlent être la source d'erreurs majeure : les résultats peuvent être significativement améliorés si ces phénomènes sont mieux pris en compte à la fois dans le lexique de prononciation et dans les modèles acoustiques. Un autre but concerne la production de transcriptions audio exactes en utilisant des corpus destinés à la presse. Même si des progrès sont à notre portée par des développements standard, une recherche plus spécifique sur la parole spontanée est nécessaire, étant donné les taux d'erreurs relativement élevés avec des modèles de langage informés.

## BIBLIOGRAPHIE


[1] M. Adda-Decker, G. Adda, J.-L. Gauvain & L. Lamel. Large vocabulary speech recognition in French. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 45-48, Phoenix, 15-19 mars 1999.

[2] C. Blanche-Benveniste. *Le français parlé, études grammaticales*, Éditions du CNRS, Paris, 1990.

[3] É. Delais-Roussarie, J. Durand. *Corpus et variation en phonologie du français : méthodes et analyses*. Presses Universitaires du Mirail, Toulouse, 2003.

[4] M.A. Morel & L. Danon-Boileau. *Grammaire de l'intonation. L'exemple du francais*, Éditions Ophrys, Paris, 1998.

[5] M. Plauche & E. Shriberg. Data-Driven Subclassification of Disfluent Repetitions Based on Prosodic Features. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 1513-1516, San Francisco, 1-7 août 1999.

[6] E. Shriberg, A. Stolcke & D. Baron. Can Prosody Aid the Automatic Processing of Multi-Party Meetings? In *Proeedings of the. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 139-146, Red Bank, 22-24 octobre 2001.

[7] S. Strassel. Simple Metadata Annotation Specification. Version 5.0, 14 mai 2003.