Cascades de transducteurs pour le chunking de la parole conversationnelle : l'utilisation de la plateforme CasSys dans le projet EPAC

Abdenour Mokrane, Nathalie Friburger, Jean-Yves Antoine

Université François Rabelais Tours – LI, IUP Blois, France
{prenom.nom}@univ-tours.fr

Résumé – Cet article présente l'utilisation de la plate-forme CasSys pour la segmentation de la parole conversationnelle (chunking) à l'aide de cascades de transducteurs Unitex. Le système que nous présentons est utilisé dans le cadre du projet ANR EPAC. Ce projet a pour objectif l'indexation et l'annotation automatique de grands flux de parole issus d'émissions télévisées ou radiophoniques. Cet article présente tout d'abord l'adaptation à ce type de données d'un système antérieur de chunking (*Romus*) qui avait été développé pour le dialogue oral homme-machine. Il décrit ensuite les principaux problèmes qui se posent à l'analyse : traitement des disfluences de l'oral spontané, mais également gestion des erreurs dues aux étapes antérieures de reconnaissance de la parole et d'étiquetage morphosyntaxique.

Abstract – This paper describes the use of the CasSys platform in order to achieve the chunking of conversational speech transcripts by means of cascades of Unitex transducers. Our system is involved in the EPAC project of the French National Agency of Research (ANR). The aim of this project is to develop robust methods for the annotation of audio/multimedia document collections which contains conversational speech sequences such as TV or radio programs. At first, this paper presents the adaptation of a former chunking system (Romus) which was developed in the restricted framework of dedicated spoken man-machine dialogue. Then, it describes the problems that are arising due to 1) spontaneous speech disfluencies and 2) errors for the previous stages of processing (automatic speech recognition and POS tagging).

Mots-clés – Traitement Automatique du Langage Parlé (TALP), segmentation, chunks, parole conversationnelle, transducteurs, Unitex.

Keywords – Spoken Language Processing, chunking, conversational speech, transducers, Unitex.

1 Introduction : le projet EPAC

Du fait du développement des technologies de l'information et de la communication, le grand public et les professionnels ont accès à une masse de données numériques de taille de plus en plus considérable. Dès lors, la question qui se pose est celle de méthodes d'accès efficaces à l'information. Elle nécessite la mise en place de techniques avancées de recherche d'information permettant une compréhension fine des requêtes et des documents manipulés. Pour fonctionner, ces techniques requièrent une indexation préalable des données qui vont être interrogées. Le projet EPAC vise la réalisation d'outils robustes d'indexation et d'annotation adaptés à un type particulier de donnée : la parole conversationnelle issue principalement de flux de données multimédias tels que les émissions radiophoniques ou télévisuelles.

Financé par l'ANR (programme MDCA - Masse de Données), EPAC réunit plusieurs laboratoires (LIUM, LIA, IRIT, LI) spécialistes du traitement de la parole et du TALN. Comme l'ont montré les campagnes d'évaluation ARPA *Broadcast News* pour l'anglais ou ESTER pour le français (Galliano *et al*. 2005), les progrès de la reconnaissance de la parole rendent possible la transcription automatique de grands flux de données audio ou multimédia. Centrés sur les journaux d'information, ces campagnes ont principalement concerné de la parole préparée ou faiblement spontanée. A l'opposé, le projet EPAC s'intéresse à des flux multimédias comprenant des séquences de parole conversationnelle caractérisée par une interactivité élevée et une forte spontanéité. Ce nouveau champ d'application nécessite la mise en œuvre d'outils spécifiques au niveau du traitement du signal : segmentation en zones de silence, parole, musique ou jingles, identification des tours de parole et des locuteurs, etc. La reconnaissance de la parole doit par ailleurs conserver sa robustesse en dépit de la nature spontanée de l'élocution et de la présence de chevauchements entre locuteurs. Les premiers résultats obtenus dans le cadre du projet montrent que la transcription automatique de la parole conversationnelle reste un objectif réaliste (Lecouteux *et al*. 2008). Au terme du projet, nous visons la diffusion d'un corpus transcrit d'une durée de 200 heures d'enregistrement.

La segmentation du flux multimédia et la transcription fournissent un corpus orthographique, accompagné de méta-données, qui est déjà utilisable par la recherche d'information. Il est toutefois intéressant de l'enrichir par différents niveaux d'annotation étudiés dans EPAC :

- Etiquetage morphosyntaxique et parenthèsage en segments minimaux des transcriptions,
- Détection et typage des entités nommées, application à l'identification du locuteur,
- Détection d'opinion pour chaque tour de parole.

Le laboratoire LI est impliqué dans la tâche de détection des entités nommées et celle de la segmentation en chunks (*chunking*) sur laquelle porte cet article. Dans un premier temps, nous présentons notre démarche qui repose sur l'application de cascades de transducteurs qui modélisent les chunks. Nous présentons ensuite la plateforme CasSys qui permet l'exécution de cascades de transducteurs Unitex. On décrit ensuite la mise en œuvre du *chunking* sur CasSys ainsi que des résultats étudiant l'influence des erreurs d'étiquetage et des disfluences.

2 Chunking de la parole conversationnelle : problèmes

La segmentation en *chunks* des énoncés s'est développée en TALN à la suite, notamment, des travaux de Steven Abney (1991). Cette notion a cependant été identifiée bien plus tôt en

linguistique et en psycholinguistique. En première approximation, on peut définir un *chunk* comme un groupe syntaxique minimal non récursif. Considérons l'énoncé suivant :

(1) [*cette petite phrase*]$_{GN}$ [*vous explicite*]$_{GV}$ [*la notion*]$_{GN}$ [*de chunk*]$_{GP}$

Il se décompose en différents chunks : nominal (GN), verbal (GV) ou prépositionnel (GP). La segmentation illustre le caractère non récursif du GP, qui n'englobe pas de GN. Suivant les approches considérées, la portée du chunk (granularité) peut cependant être variable. Dans tous les cas, le parenthésage en chunks revêt plusieurs intérêts pour la recherche d'information dans un flux de parole conversationnelle :

- Centré sur un mot lexical unique, le chunk correspond à une unité minimale de sens dans l'univers du discours, sur laquelle peut reposer la recherche d'information. Ainsi, les entités nommées correspondent toujours à un ou plusieurs chunks.

- Le chunk est adapté à la parole conversationnelle. Il est en effet le lieu de réalisation privilégié de l'entassement paradigmatique (Blanche-Benveniste, 1997:47) mis en jeu par les répétitions ou les réparations de l'oral spontané. Lorsqu'une disfluence a une portée supérieure au chunk, elle n'affecte en aucune manière les chunks extérieurs.

Cette dernière remarque est importante, car les disfluences orales (hésitations, répétitions, réparations, incises) cassent souvent la structure syntaxique des énoncés, ce qui rend d'autant plus difficile leur analyse automatique.

De nombreux systèmes de chunking efficaces (Giguet, Vergne 1997) ont été développés pour le langage écrit. Les disfluences de l'oral spontané interdisent toutefois leur application directe à la parole hautement conversationnelle. C'est ainsi que la communauté parole s'est tournée, à la suite de (Hindle 1983), vers des approches plus ad-hoc de pré-correction avant analyse: on détecte des patterns assez simples de reprises, dont le *reparandum* est ensuite effacé pour normaliser l'énoncé (Bear *et al*. 1992, Heeman & Allen 2001). Ces techniques ont donné de bons résultats en détection, mais opèrent parfois des effacements abusifs. Plus globalement, l'effacement du reparandum peut gommer une information utile. Considérons les deux exemples ci-dessous :

(2) *Je cherche* [*un camping près de la gare*]$_{REP}$ [*euh non*]$_{ED}$ *un près de la côte pardon*

(3) *Barton Fink est* [*un film dense*]$_{REP}$ *un film porté par un scénario foisonnant*

Dans l'exemple (2), la suppression du reparandum avant la zone d'édition *euh non* empêche le calcul de la référence dans l'altération *un près de la côte*. L'exemple (3) correspond à une répétition avec enrichissement lexical. Effacer le reparandum revient à perdre une information qui n'est en rien corrigée par l'altération qui suit. Les conséquences de ces effacements abusifs peuvent être importantes. Aussi adoptons nous une démarche non destructive.

3 Chunking incrémental de la parole : cascade de transducteurs

Nous proposons une analyse incrémentale fondée sur la détection d'ilots de certitude : les chunks non affectés par les disfluences. Dans un premier temps, on applique des règles de segmentation décrivant les structures « légitimes » des chunks. Les zones non segmentées à l'issue de cette étape sont marquées comme disfluentes. Il est alors envisageable d'appliquer des règles spécifiques pour caractériser les différentes parties des disfluences (reparandum,

zone d'édition) sans les effacer. Cette démarche rejoint les principes du TAL robuste, à savoir (Aït-Mokhtar *et al*. 2003) que l'analyse est complète mais superficielle (*shallow parsing*), qu'elle est non destructrice (on conserve l'information pour les étapes ultérieures) et suit une stratégie incrémentale où chaque niveau utilise une connaissance qui fait sens par elle-même (indépendance conceptuelle).

Nous avions adopté cette approche dans le système ROMUS de compréhension automatique de la parole (Goulian *et al*. 2003 ; Antoine *et al*. 2003). Dans ROMUS, la structure des chunks est décrite par des expressions régulières travaillant sur les parties du discours associées aux mots. Ces expressions sont compilées en transducteurs déterministes à l'aide du toolkit FSA (Van Noord 1997). Chaque transducteur est utilisé en cascade pour introduire dans l'énoncé des marqueurs de délimitation, jusqu'à arriver à une segmentation complète. L'ambiguïté est gérée par une heuristique de maximisation des segments construits. Ces principes sont repris par le système SECARE que nous avons réalisé pour EPAC, avec trois particularités supplémentaires :

- le champ d'application de SECARE n'est plus le dialogue oral homme-machine finalisé mais la langue générale. Il n'est donc plus possible de s'appuyer sur une connaissance pragmatique pour résoudre la caractérisation finale des zones disfluentes,
- alors que ROMUS travaillait sur de la parole transcrite, SECARE opérera sur les sorties réelles de la reconnaissance de la parole, fournies par les laboratoires LIA ou LIUM,
- Les cascades de transducteurs ne sont plus implantées sur le toolkit FSa, mais sur la plateforme CasSys/Unitex. L'intérêt d'Unitex est de fournir une représentation explicite de la structure des chunks, qui peut donc être manipulée par des linguistes non informaticiens.

4 CasSys / Unitex

CasSys est un système de cascade de transducteurs, développé au LI, utilisant des outils proposés par Unitex (Paumier 2003). La cascade est une suite de transducteurs, au format Unitex, passés dans un ordre précis afin par exemple d'extraire ou de remplacer des motifs, ou encore d'enrichir le texte avec un balisage XML (Friburger. 2002).

Sous Unitex, les transducteurs sont représentés par des graphes (figure 1) facilement lisibles. Le fonctionnement interne "simplifié" d'Unitex est le suivant. Unitex mémorise dans un fichier tous les motifs localisés et leur emplacement dans le texte analysé. Puis si on demande une concordance, Unitex transforme le texte en fonction du mode choisi : en mode remplacement, il remplace les entrées du transducteur par ses sorties, alors qu'en mode fusion, il fusionne les entrées et sorties.

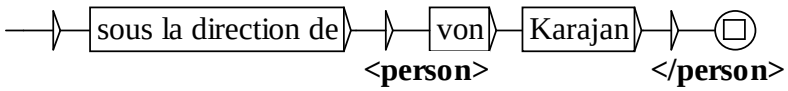

Figure 1: Un exemple de transducteur Unitex

CasSys ajoute à Unitex la possibilité d'extraire un motif du texte (ce motif étant enrichi des éventuelles sorties du transducteur) pour le mémoriser dans un fichier tandis qu'on le remplace dans le texte par une étiquette qui permettra de le retrouver plus tard. Suivant les principes de l'analyse en cascade, les transducteurs sont donc passés de sorte qu'on reconnaisse en premier

les motifs les moins ambigus; ceux ci sont supprimés du texte et par la suite ne risquent pas d'être confondus avec un motif reconnu par un autre transducteur.

A titre d'illustration, le graphe représenté dans la figure 1 reconnaît exclusivement la phrase « *sous la direction de von Karajan* ». On applique ce transducteur sur le texte suivant (4) pour obtenir après application la séquence (4'):

(4) *Le concert a lieu, sous la direction de von Karajan, en Bavière.*
(4') *Le concert a lieu, sous la direction de <$exemple1:0$>, en Bavière.*

Les entrées reconnues et les sorties du transducteur sont fusionnées pour donner la séquence *<person> von Karajan </person>* qui est extraite du texte et placée dans un index à la position 0. On peut choisir que seule la partie reconnue entre les balises *<person>* et *</person>* soit extraite du texte, dans ce cas la séquence *sous la direction de* est toujours dans le texte. L'étiquette *<$exemple1:0$>* insérée dans le texte indique quel est le graphe qui a reconnu cette séquence (exemple1) et permet de retrouver la séquence correspondante dans le fichier index (position 0). A la fin de la cascade, les motifs extraits sont replacés dans le texte:

(4'') *Le concert a lieu, sous la direction de <person> von Karajan </person>, en Bavière.*

5 Implantation des cascades de transducteurs

5.1 Formats de données : segmentation PEAS

Un des résultats attendus du projet EPAC est la mise à disposition d'un grand corpus de parole conversationnelle transcrite et annotée. La réutilisabilité de cette ressource est favorisée par la normalisation des formats d'encodage, qui utilisent tous XML. Les transcriptions suivent le format *.trs* Transcriber. Afin d'ajouter des couches d'annotations indépendantes sur les transcriptions, nous avons défini une référence temporelle de synchronisation qui se base sur une segmentation des transcriptions en tokens. Cette référence s'inspire du format utilisé dans le projet européen LUNA (www.ist-luna.eu/).

La segmentation en chunks repose sur le format PEAS utilisé lors de la campagne de test EASy (Paroubek *et al*. 2006). Le choix de PEAS relève de notre volonté de normalisation. Il peut être considéré en effet comme un format d'échange accepté par l'ensemble de la communauté francophone. Issu de difficiles compromis, PEAS a conduit à une simplification extrême de la portée des chunks. Par exemple, la séquence de mots *très très haut* y est annotée comme la suite de deux groupes adverbiaux (GR) suivi d'un groupe adjectival (GA) alors qu'il est clair que les adverbes de degrés dépendent de l'adjectif qu'ils qualifient. PEAS rend compte de ces dépendances par des relations entre chunks. Il nous semble toutefois regrettable d'identifier des dépendances aussi locales et, par exemple, la relation de sous-catégorisation entre un prédicat verbal et ses arguments. Nous ne nous interdisons pas de regrouper en interne certains types de chunks pour accroître leur portée. Mais les données qui seront diffusées suivront la norme PEAS, à laquelle nous avons apporté deux compléments :

- Catégories spécifiques aux disfluences orales, qui avaient pas été étudiées dans EASy et donc par PEAS, qui est plus centré sur l'écrit : REP (reparandum) et ED (zone d'édition),
- Catégories spécifiques pour assurer un parenthésage complet de l'énoncé. Par exemple, ajout d'un chunk COO pour représenter les conjonctions de coordination. La coordination

est représentée dans PEAS par une relation de dépendance. Il nous semble plus justifié de lui accorder le statut de chunk, à la fois pour atteindre une segmentation complète, mais également parce que les coordinations peuvent contenir des disfluences orales complexes.

Pour rappel, PEAS distingue à la base les chunks suivants: NV (noyau verbal), PV (groupe verbal infinitif introduit par une préposition), GN (groupe nominal), GP (groupe prépositionnel), GA (groupe adjectival sans les adjectifs antéposés) et GR (groupe adverbial).

5.2 Implantation des cascades sur CasSys / Unitex

Comme nous l'avons vu (cf. § 3), la segmentation est basée sur une cascade de transducteurs qui identifie dans une première passe les chunks qui ont une structure normée, suivant une stratégie par ilots de confiance. Ce n'est que dans un second temps que l'on s'intéresse aux zones non identifiées, afin de réaliser une segmentation complète des énoncés. A l'heure actuelle, seules des transcriptions manuelles de parole conversationnelle ont été diffusées aux partenaires du projet EPAC. Dans l'attente de la réception de transcriptions automatiques, et compte tenu de l'influence centrale des erreurs de reconnaissance sur la robustesse des systèmes, nous avons choisi de n'implémenter intégralement que la première cascade et d'en étudier les limites. La seconde passe se limite à la caractérisation des catégories complémentaires aux annotations PEAS, tel que le chunk COO et le chunk PONCT (pour la ponctuation). Elle attribue enfin l'étiquette CHINC (chunk inconnu) aux zones non encore segmentées. Ces séquences seront par la suite analysées, soit pour caractériser les disfluences, soit pour corriger des erreurs de reconnaissance ou d'étiquetage morphosyntaxique. La chaîne globale de traitement est illustrée dans la figure 2.

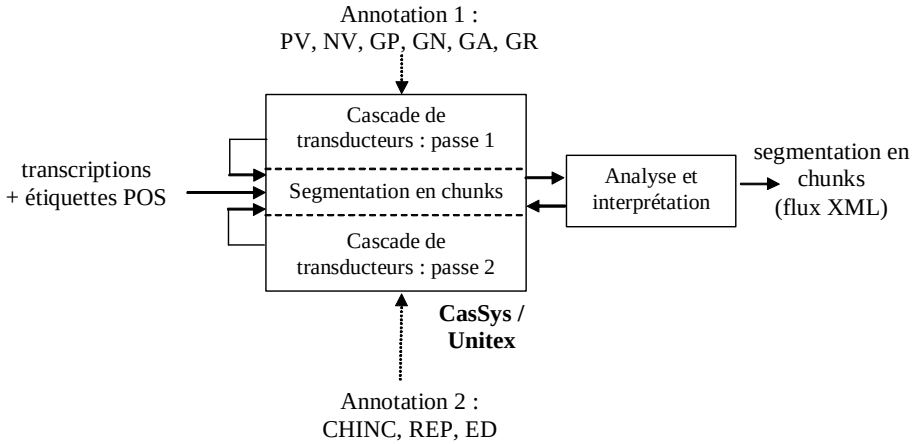

Figure 2: chaîne de traitements pour le chunking

L'analyse repose sur la définition d'un transducteur par type de chunk (GN, NV, etc.). A chaque chunk est associé un transducteur principal qui décrit sa structure syntaxique et une série de transducteurs intermédiaires dédiés à la reconnaissance des mots avec leurs tags et à la gestion des flux XML. Au final, la séquence de mots étiquetés ci-dessous :

```
<word id="s0034_w0001" token="s0034_t0005" pos="AINDMP"> tous_les </word>
<word id="s0034_w0002" token="s0034_t0007" pos="NMP"> jours </word>
```

permet la génération, via la plate forme CasSys, du chunk GN suivant, également en XML :

```
<chunk token_deb="s0034_t0005" word_deb="s0034_w0001"
token_fin="s0034_t0007" word_fin="s0034_w0002" id="s0034_c"> GN </chunk>
```

Nous avons également ré-implémenté certains automates prédéfinis sous Unitex, tel que l'automate <MOT>, ceci afin de supporter certaines disfluences (amorces de mots inachevés, par exemple). La base d'identification des chunks de la première passe est composée de 386 transducteurs. La figure 3 donne en exemple de transducteur principal associé au chunk GN.

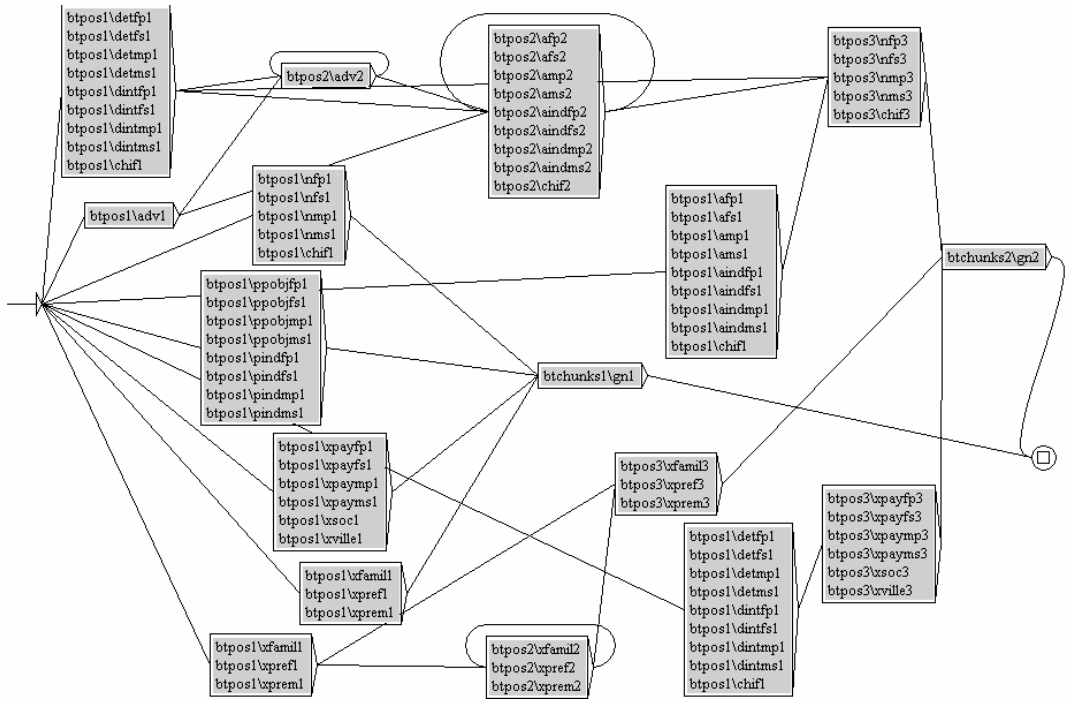

Figure 3: Transducteur principal associé au chunk GN

L'ordre d'application des transducteurs dans la cascade est essentiel, puisqu'il permet de gérer les ambiguïtés d'analyse, notamment pour les chunks qui se recouvrent. On comprend ainsi aisément que le transducteur GP doit être appliqué avant celui du GN. Un transducteur intermédiaire GNpourGP permet par contre d'appeler la recherche de motifs de type GN une fois passée la préposition (figure 4). Il en va de même pour PV par rapport à NV. Au final, la première cascade suit l'ordre PV puis NV, GP, GN, GA, et GR. L'application de NV avant GN est rendue nécessaire par l'inclusion potentielle de GN pronominaux dans les noyaux verbaux (pronoms personnels sujets ou clitiques). De même, les GN peuvent inclure des adjectifs antéposés, GA doit donc être appliqué après GN dans la cascade. Dans la seconde passe, l'ordre est le suivant: PONCT, COO, puis CHINC. Appliqué en dernier, le transducteur CHINC (segments inconnus) servira ultérieurement à l'analyse fine des disfluences.

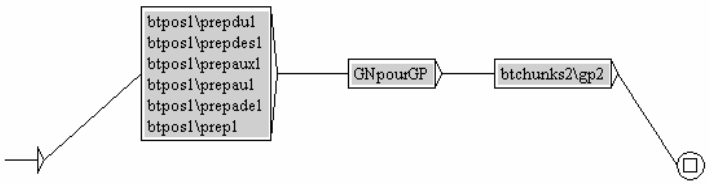

Figure 4: Exemple de cas de chunks en recouvrement : transducteur GP

A titre d'exemple, le chunk CHINC ci-dessous est dû à une non reconnaissance (MOTINC : mot hors vocabulaire) de noms propres par le tagueur LIA_TAGG[1] du laboratoire LIA. Nous revenons sur ce problème dans le paragraphe suivant.

---

1 LIA_TAGG : http://old.lia.univ-avignon.fr/chercheurs/bechet/download_fred.html

```
<word id="s0002_w0003" token="s0002_t0005" pos="MOTINC"> Miroslav </word>
<word id="s0002_w0004" token="s0002_t0007" pos="MOTINC"> Marcelli </word>
<chunk token_deb="s0002_t0005" word_deb="s0002_w0003"
token_fin="s0002_t0007" word_fin="s0002_w0004" id="s0002_c">CHINC</chunk>
```

6 Résultats

Le système SECARE a été évalué sur les transcriptions manuelles déjà réalisées dans le cadre d'EPAC. Il n'est pas inutile de rappeler que ni EASy ni la première campagne d'évaluation ESTER ne se sont intéressées au chunking de la parole conversationnelle. Il n'existe donc pas de résultats de référence en la matière. Les tests ont été effectués sur un extrait d'émission radiophonique à forte interactivité, regroupant 893 chunks. Les transcriptions ont été annotées en partie du discours par l'étiqueteur LIA_TAGG. Ces annotations comportent un certain taux d'erreurs, puisque LIA_TAGG n'a pas encore été adapté à la tâche EPAC.

SECARE a un comportement robuste en présence de disfluences. Les réparations ne créent pas de faux positifs et n'induisent pas d'erreurs dans la délimitation des chunks réguliers. Comme on pouvait s'y attendre, les erreurs de tagging ont par contre une influence directe sur les performances du système. Ainsi, les mots inconnus de LIA_TAGG peuvent fausser la délimitation des chunks. Considérons l'exemple suivant :

(5) sortie LIA_TAGG) [*le*]$_{\text{DETMS}}$ [*livre*]$_{\text{NMS}}$ [*de*]$_{\text{PREP\_ADE}}$ [*Pierre*]$_{\text{XPREM}}$ [*Péan*]$_{\text{MOTINC}}$
segmentation [*le livre*]$_{\text{GN}}$ [*de Pierre*]$_{\text{GP}}$ [*Péan*]$_{\text{CHINC}}$

LIA_TAGG ne connait pas le patronyme *Péan* qui est étiqueté comme mot inconnu. SECARE ne peut alors identifier le rattachement de ce dernier au groupe prépositionnel. Cette situation est particulièrement pénalisante dans le cas des entités nommées.

Dans l'exemple ci-dessous (6), ce n'est pas la répétition du déterminant «*une*» qui trompe le système, mais le fait que sa seconde occurrence est étiquetée par erreur comme un adjectif :

(6) sortie LIA_TAGG [*d'*]$_{\text{PREP\_ADE}}$ [*une*]$_{\text{DETFS}}$ [*une*]$_{\text{AFS}}$ [*génération*]$_{\text{NFS}}$
segmentation [*d'une une génération*]$_{\text{GP}}$

Au final, on récupère un faux positif équivalent structurellement au chunk «*d'une belle génération*» alors que SECARE aurait normalement bien détecté la présence d'une disfluence.

Ces observations se retrouvent d'un point de vue quantitatif. Nous avons utilisé plusieurs métriques de test, afin de caractériser aussi bien les erreurs de typage que de délimitation :

- Rappel (R), précision (P) et F-score de la segmentation,
- Les taux d'insertion (I), suppression (D) et substitution (S) quantifient mieux les erreurs de délimitation de chunks. Si un GP attendu est scindé en un GR et un GN, trois erreurs vont être imputées au système: la suppression du GP et l'insertion du GR et du GN.

Le tableau 1 présente les performances de SECARE sur l'ensemble du corpus et sur les chunks dans lesquels le LIA_TAGG n'a pas fait d'erreur d'étiquetage. Pour information, 9,4% des chunks étudiés présentaient au moins un mot avec une partie du discours erronée. Ces résultats suggèrent que nos transducteurs sont robustes. Si le F-score n'est globalement que de 0.805, les segmentations erronées sont majoritairement dues aux erreurs d'étiquetage morphosyntaxique. En effet, le F-score approche 0.94 sur les chunks sans erreur d'étiquetage.

La plupart des erreurs du système correspond au découpage des chunks attendus en plusieurs chunks différents. En particulier, si un mot a été mal étiqueté par LIA_TAGG, il est fréquent que le chunk correct soit divisé en deux ou trois chunks erronés, ce que traduit le fort taux d'insertions. A l'opposé, il est extrêmement rare (3 observations sur l'ensemble du corpus de test) que les frontières du chunk attendu ne se retrouvent pas dans la segmentation erronée.

Corpus	R	P	F-score	I	D	S
Intégral (893 chunks)	85,1%	76,3%	0.805	20,0 %	10,8 %	3,7%
Sans erreurs de tagging (816 chunks)	95.3%	92.6%	0.939	n.c.	n.c.	n.c.

Tableau 1: Performances du système SECARE sur un corpus de transcriptions manuelles.

Il est enfin à remarquer qu'une partie assez significative des erreurs de LIA_TAGG est due à la présence de mots inconnus dans les entités nommées (patronymes, toponymes, etc.). La fréquence de ces erreurs baissera sensiblement lorsque le LIA livrera un étiqueteur adapté à la tâche EPAC. Mais on peut déjà remarquer qu'une part non négligeable de ces erreurs est assez facilement modélisable. Par exemple, un mot inconnu avec majuscule initiale précédé d'un prénom a toutes les chances d'être un patronyme. Un travail assez rapide d'ajout de transducteurs de post-correction dans la seconde cascade nous a ainsi permis d'atteindre un F-score de 0.884 (rappel : 90,7% ; précision 86, 2%) sur le corpus de test complet.

On peut être étonné par notre stratégie d'analyse séquentielle (étiquetage puis segmentation) qui ne peut que cumuler les erreurs, alors qu'on sait que l'étiquetage morphosyntaxique gagne à être conduit en parallèle avec le chunking (Giguet, Vergne 1997). Cette observation demanderait à être confirmée sur de la parole conversationnelle. Là n'est pas toutefois la justification de notre approche, qui découle en fait des objectifs du projet EPAC. Celui-ci vise à évaluer le gain d'une révision manuelle d'annotations automatiques par rapport à une annotation purement manuelle. A terme, SECARE travaillera donc sur des données révisées sans erreurs. Pour d'autres applications, nous envisageons par contre de coupler annotation et segmentation en utilisant le dictionnaire Delas, fourni avec Unitex, étendu par la base Prolex de noms propres (Tran, Maurel 2006).

7 Conclusion et perspectives

Les performances de SECARE montrent qu'il est possible de généraliser à la langue générale les techniques de segmentation que nous avions développées pour le dialogue homme-machine. Il reste toutefois à évaluer le système sur des transcriptions automatiques pour confirmer cette observation. Dans l'immédiat, ces premières expériences montrent que le système est robuste sur des transcriptions exactes de parole spontanée. Nous allons maintenant compléter la seconde cascade de transducteurs pour distinguer, parmi les zones inconnues, les segments disfluents (reparandum et zone d'édition).

Remerciements

Ce projet est financé par l'Agence Nationale de la Recherche (projet ANR-06-MDCA-2006). Nous remercions Denis Maurel pour son aide sur l'utilisation du système Unitex.

Références

ABNEY S. (1991) Parsing by chunks, In. Berwick, Abney, Tenny (Eds.) *Principle-based parsing. Amsterdam*. Kluwer Academic Publ. Dordrecht, Pays-Bas.

AÏT-MOKHTAR S., CHANOD J.-P., ROUX C. (2003) Robustness beyond shallowness: incremental deep parsing, *Natural Language Enginerring*, Vol. 8 (3-2).

ANTOINE J.-Y.., GOULIAN J., VILLANEAU J. (2003) Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la comprehension de la parole spontanée. *TALN'2003*. Batz.

BEAR J., DOWDING J., SHRIBERG E. (1992) Integrating multiple knowledge sources for detection and correction of repairs in Human-Computer dialogue, Proc. *Annual meeting of the ACL, ACL'92*, Newark, Danemark. pp. 56-63.

BLANCHE-BENVENISTE C. (1997) Approches de la langue parlée en français, Coll. L'essentiel Français, Ophrys, Paris, France.

FRIBURGER N. (2002) Reconnaissance automatique des noms propres; application à la classification automatique de textes journalistiques. Thèse de doctorat, U. Fr. Rabelais Tours.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F., GRAVIER G.(2005) The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News, Actes *Eurospeech/Interspeech'2005*, Lisbonne, Portugal.

GIGUET E., VERGNE J. (1997) From Part-of-Speech Tagging to Memory-based Deep Syntactic Analysis. Proc. IWPT'97, MIT, Boston, Massachussets, USA.

GOULIAN J., ANTOINE J.-Y., POIRIER F. (2003) How NLP techniques can improve speech understanding Actes *Eurospeech'2003*, Genève, Suisse. 2773-2776.

HEEMAN P., ALLEN J. (2001) Improving robustness by modelling spontaneous speech events, In. Robustness in language and speech technology, Kluwer, Dordrecht, Pays-Bas, pp. 123-152.

HINDLE D. (1983) Deterministic parsing of syntactic nonfluencies. Actes *ACL'83*, pp. 123-128

TRAN M., MAUREL D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, Traitement automatique des langues, Vol. 47(3), 115-139

LECOUTEUX B., LINARÈS G., ESTÈVE Y., GRAVIER G. (2008), Generalized driven decoding for speech recognition system combination, Actes *IEEE ICASSP 2008*, Las Vegas, Nevada,USA.

PAROUBEK P., ROBBA I.,VILNAT A. AYACHE C. (2006) Data Annotations and Measures in EASY the Evaluation Campaign for Parsers of French, Actes *5th international Conference on Language Resources and Evaluation, LREC 2006*, Gênes, Italie, pp.315-320.

PAUMIER S. (2003) De la reconnaissance de formes linguistiques à l'analyse syntaxique, Thèse de Doctorat, Université de Marne-la-Vallée, France.

VAN NOORD G. (1997) Fsa utilities: a toolbox to manipulate finite state automata. In Raymond D. *et al.* (Eds.) *Automata Implementation*, Springer Verlag, RFA. pp. 87-108.