Traitement des disfluences dans le cadre de la compréhension automatique de l'oral arabe spontané

Younès Bahou, Abir Masmoudi, Lamia Hadrich Belguith

ANLP Research Group – Laboratoire MIRACL
Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088, 3018 - Sfax – TUNISIE
Téléphone (216) 74 278 777, Fax (216) 74 279 139
bahou_younes@yahoo.fr, masmoudiabir@gmail.com, l.belguith@fsegs.rnu.tn

Résumé. Les disfluences inhérents de toute parole spontanée sont un vrai défi pour les systèmes de compréhension de la parole. Ainsi, nous proposons dans cet article, une méthode originale pour le traitement des disfluences (plus précisément, les autocorrections, les répétitions, les hésitations et les amorces) dans le cadre de la compréhension automatique de l'oral arabe spontané. Notre méthode est basée sur une analyse à la fois robuste et partielle, des énoncés oraux arabes. L'idée consiste à combiner une technique de reconnaissance de patrons avec une analyse sémantique superficielle par segments conceptuels. Cette méthode a été testée à travers le module de compréhension du système SARF, un serveur vocal interactif offrant des renseignements sur le transport ferroviaire tunisien (Bahou et al., 2008). Les résultats d'évaluation de ce module montrent que la méthode proposée est très prometteuse. En effet, les mesures de rappel, de précision et de *F-Measure* sont respectivement de 79.23%, 74.09% et 76.57%.

Abstract. The disfluencies inherent in spontaneous speaking are a real challenge for speech understanding systems. Thus, we propose in this paper, an original method for processing disfluencies (more precisely, self-corrections, repetitions, hesitations and word-fragments) in the context of automatic spontaneous Arabic speech understanding. Our method is based on a robust and partial analysis of Arabic oral utterances. The main idea is to combine a pattern matching technique and a surface semantic analysis with conceptual segments. This method has been evaluated through the understanding module of SARF system, an interactive vocal server offering Tunisian railway information (Bahou et al., 2008). The evaluation results of this module show that the proposed method is very promising. Indeed, the measures of recall, precision and F-Measure are respectively 79.23%, 74.09% and 76.57%.

Mots-clés : disfluences, segment conceptuel, reconnaissance de patrons, parole arabe spontanée.

Keywords: disfluencies, conceptual segment, pattern matching, spontaneous Arabic speech.

1 Introduction

Dans ce papier, nous proposons une méthode originale pour le traitement des disfluences (en particulier, les autocorrections, les répétitions, les hésitations et les amorces) dans le cadre de la compréhension automatique de la parole arabe spontanée. Cette méthode est basée sur une analyse robuste et partielle par segments conceptuels des énoncés oraux arabes. L'originalité de la méthode réside dans l'utilisation à la fois de segments conceptuels et d'une technique de reconnaissance de patrons (*pattern matching*) pour le traitement des disfluences. Ainsi, un énoncé étiqueté sémantiquement passe par trois étapes de traitement à savoir, une étape de découpage en segments conceptuels, une étape de délimitation des segments disfluents contenant les disfluences et une étape de détection et de correction des disfluences basée sur la technique de la reconnaissance de patrons.

Ce travail entre dans le cadre de la réalisation du serveur vocal interactif SARF (Bahou et al., 2008 ; Hadrich Belguith et al., 2009) offrant des renseignements sur le transport ferroviaire tunisien en langue arabe standard moderne (*e.g.,* horaire d'un train, tarification, etc.). Nous visons, à travers le présent travail, à appliquer et à tester notre méthode de traitement des disfluences dans l'oral arabe et ce à travers le module de compréhension du système SARF.

Cet article s'articule autour de cinq sections principales. La section 2 explique les principaux types de disfluences. La section 3 expose un aperçu sur les approches de traitement des disfluences. La section 4 détaille la méthode que nous proposons pour le traitement des disfluences de l'arabe parlé et la section 5 présente la mise en œuvre et l'évaluation de la méthode proposée ainsi qu'une discussion des résultats obtenus.

2 Les disfluences dans l'oral spontané

Comme indique leur étymologie, les disfluences correspondent à toute interruption ou perturbation de la *fluence*, c'est-à-dire du cours normal de la production orale spontanée (Bove, 2008). Dans ce qui suit, nous mettons l'accent sur les principaux phénomènes de disfluences à savoir, les autocorrections, les répétitions, les hésitations et les amorces.

- **Autocorrection :** il s'agit du cas où le locuteur fait une ou plusieurs erreurs et se corrige dans le même énoncé. Dans ce cas le mot (ou les mots) erroné est prononcé complètement (Bouraoui, 2009).

- **Répétition :** il s'agit du cas de la répétition d'un mot ou d'une série de mots. Elle est définie sur des critères purement morphologiques (Kurdi, 2003).

- **Hésitation :** il s'agit du cas d'une pause remplie dans la production orale qui peut se manifester de diverses manières : soit par le recours à un morphème spécifique (*e.g., euh, hum,* etc.), soit en prenant la forme d'un allongement de syllabe (Bove, 2008).

- **Amorce :** il s'agit du cas de l'arrêt de la production d'un mot avant la fin normale de celui-ci. Dans sa terminologie, une amorce correspond toujours à un fragment de mot qui peut être identifié grâce à la connaissance de la phraséologie (Bouraoui, 2009).

3 Travaux antérieurs

Dans la littérature, les disfluences sont traitées soit au niveau de la reconnaissance vocale soit au niveau de la compréhension de la parole. Ainsi, nous avons répertorié les travaux existants selon ces deux niveaux de traitement à savoir, niveau reconnaissance et niveau compréhension.

3.1 Niveau reconnaissance

La reconnaissance automatique de la parole consiste à extraire la liste des mots contenue dans un signal vocal. À ce niveau de traitement et à notre connaissance, il y a une seule approche qui a été proposée pour le traitement des disfluences à savoir, l'approche de SRI proposée par Bear et al. au sein de *Standford Research Instute* (SRI) (Bear et al., 1992).

Le travail de Bear et al. consiste, en première étape, à proposer un schème d'annotation des disfluences qui combine la simplicité à la finesse nécessaire pour la représentation des différentes formes de disfluences (Bear et al., 1992). Il s'agit, ensuite, de combiner l'analyse syntaxique et sémantique (afin de réduire la surgénération de patrons) avec la technique de la reconnaissance de patrons. Le but étant de détecter et de corriger les répétitions simples et les anomalies syntaxiques simples comme « a the » (Bear et al., 1992). L'inconvénient de cette combinaison est qu'elle est incompatible avec les approches d'analyse partielle qui sont les plus adaptées au traitement de l'oral. Ainsi, l'approche de SRI rend le module de traitement des disfluences complètement dépendant de l'analyseur syntaxique et par conséquent elle réduit considérablement sa portabilité.

Heeman et Allen ont adopté cette approche dans le cadre du projet TRAINS au sein de l'université de *Rochester* (Heeman, Allen, 1996). La première étape de ce travail consiste à proposer une version modifiée du schème d'annotation des chercheurs de SRI. Ainsi, le schème proposé ne permet pas le partage de la zone remplacée dans le cas de disfluences imbriquées. L'idée principale de ce travail est basée sur l'usage de certains types de disfluences comme indices des occurrences des autocorrections. Pour cela, les auteurs mettent en place un modèle statistique (n-grammes) permettant d'associer à la probabilité d'occurrence d'une disfluence donnée, la probabilité d'apparition d'une autocorrection. Il s'agit ensuite de détecter et de corriger les disfluences en utilisant des patrons et des règles. L'inconvénient principal du travail de Heeman et Allen est que l'utilisation des tags comme source unique de connaissance morphologique est trop limitative. En effet, dans certains cas on a besoin d'informations morphologiques détaillées (*e.g.,* personne, fonction syntaxique, etc.) afin de pouvoir analyser correctement un cas de disfluences.

3.2 Niveau compréhension

La compréhension se situe en aval de la reconnaissance vocale. Son rôle est d'interpréter sémantiquement et d'attribuer une représentation sémantique de l'ensemble d'éléments lexicaux (habituellement un ou plusieurs énoncés) qui ont été produits par le système de reconnaissance. À ce niveau de traitement, nous distinguons deux principales approches pour le traitement des disfluences à savoir, l'approche à base de patrons et l'approche à base de méta-règles syntaxiques.

- **Approche à base de patrons**

L'avantage de cette approche est la simplicité de sa mise en œuvre, puisque les algorithmes de détection des disfluences à base de patrons sont relativement simples à concevoir. De plus, elle est généralisable à plusieurs types de corpus, de tâches et aussi de langues.

Le travail de Kurdi porte sur le traitement du langage oral spontané dans le contexte du dialogue oral Homme-machine à travers le système CORRECTOR (Kurdi, 2003). Dans son travail, Kurdi a désigné les disfluences par le terme « extragrammaticalités ». Ainsi, l'auteur tente de mêler deux tendances : ***i)*** la première tendance consiste à utiliser les deux techniques de n-grammes et de la reconnaissance de patrons pour traiter certains phénomènes comme les répétitions et les autocorrections, ***ii)*** la deuxième tendance consiste à modéliser syntaxiquement les amorces et les inachèvements.

Le travail de Bove s'inscrit dans le cadre d'un travail d'équipe sur l'analyse morphologique et syntaxique du français parlé (Bove, 2008). Il adopte une méthode « hybride » de reconnaissance de patrons (basée

sur les catégories morpho-syntaxiques du corpus) couplée à un calcul de n-grammes pour détecter les différents phénomènes de disfluences. Ensuite, les énoncés disfluents précédemment analysés seront découpés en syntagmes minimaux non récursifs (ou *chunks*). Le corpus final est ainsi segmenté en *chunks* habituels (qu'on retrouve lors de l'analyse de l'écrit) à-côté des *chunks* disfluents dégagés lors de l'analyse (Bove, 2008). Dans son travail, l'auteur n'a pas proposé une méthode pour la correction des disfluences détectées.

L'approche à base de patrons présente certaines limites. Parmi ces limites, la stabilité des patrons obtenus. Le terme de stabilité désigne le nombre de patrons différents pouvant être dérivés à partir d'un phénomène donné. Dans la plupart des études sur les disfluences, un très grand nombre de patrons différents est obtenu pour représenter les disfluences, ou même un seul type de disfluences, telles que les autocorrections. Ainsi, la prise en compte extensive de cette multiplicité de patrons implique un algorithme complexe et difficile à mettre en œuvre.

- **Approche à base de méta-règles syntaxiques**

Cette approche est proposée par Core et Schubert dans le cadre de l'analyse robuste des dialogues au sein du groupe de dialogue de l'université de *Rochester* (Core, Schubert, 1999). La particularité principale de ce travail est l'introduction des informations linguistiques (notamment syntaxiques) dans le traitement des disfluences, d'une manière originale.

Dans le travail de Core et Schubert, le traitement de disfluences se fait en deux étapes. La première étape consiste à détecter les disfluences à l'aide d'un modèle de langage statistique. Par ailleurs, la fonction principale de ce module est de détecter les disfluences et de proposer une première délimitation de chacune d'entre elles. Alors que la deuxième étape consiste à donner une interprétation couvrant la totalité des mots de l'énoncé d'entrée. Pour cela, les disfluences détectées par le module statistique sont analysées à l'aide de méta-règles dédiées spécialement à cette tâche (Core, Schubert, 1999).

Le travail de Mckelvie (Mckelvie, 1998) adopte cette approche dans le cadre d'un projet qui vise l'analyse syntaxique des dialogues oraux spontanés au sein de l'université d'*Edinburgh*. Comparé au travail de Core et Schubert, Mckelvie ajoute deux catégories syntaxiques outres celles qui sont classiques à savoir, les syntagmes d'éditions (spécifiques aux disfluences) et les marqueurs discursifs (*e.g.,* oui, bon, etc.). L'idée principale de ce travail reste semblable à celle de Core et Schubert. Le plus apporté par l'auteur est le fait d'ignorer tous les syntagmes d'éditions et les marqueurs discursifs qui apparaissent après un constituant de confiance (pertinent et correct pour l'auteur). Ainsi, l'analyseur syntaxique proposé pour le traitement des dialogues oraux garanti une certaine robustesse face aux disfluences.

L'inconvénient principal de cette approche réside dans l'utilisation des méta-règles syntaxiques. D'un côté, les méta-règles syntaxiques sont difficiles à recenser et à mettre en œuvre. D'un autre côté, l'utilisation de méta-régles syntaxiques dans une analyse multiplie le temps de calcul d'environ trois fois.

4 Méthode proposée pour le traitement des disfluences

Avant d'exposer la méthode que nous proposons, nous jugeons nécessaire de présenter le corpus d'étude que nous avons utilisé.

4.1 Corpus d'étude

Vu que les ressources linguistiques arabes sont très rares, nous étions amenés à créer notre propre corpus d'étude constitué de 300 dialogues (soit 11 heures d'enregistrements) selon la technique de *Magicien d'Oz*. Cette technique fait simuler le comportement de la machine par un *compère* humain (*i.e.,* le magicien) à l'insu de l'utilisateur. Ce corpus nous a permis d'étudier quantitativement et qualitativement les différents types de disfluences que nous envisageons de traiter.

Par l'étude quantitative, nous visons à étudier les occurrences d'apparition des différents types de disfluences dans le corpus d'étude. Parmi les types de disfluences, nous nous sommes intéressés aux quatre types suivants : les autocorrections, les répétitions, les hésitations et les amorces vu leur fréquence dans les dialogues oraux arabes. Notons que cette étude nous est utile pour comparer les caractéristiques de l'oral arabe spontané par rapport à d'autres langues, précisément le français. La figure 1 résume les résultats obtenus.

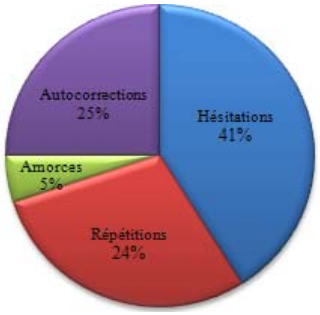

Figure 1 : Distribution des types de disfluences dans le corpus d'étude

À la lumière de cette distribution et en comparaison avec l'étude menée par Bove (Bove, 2008) sur le français parlé, nous avons constaté que les autocorrections dans l'oral arabe spontané sont plus fréquentes que celles dans l'oral français spontané (*i.e.,* Bove a trouvé 13% d'autocorrections dans son corpus d'étude) alors que les répétitions et les amorces sont moins fréquentes (*i.e.,* Bove a trouvé respectivement 44% et 15%).

Par l'étude qualitative, nous visons à recenser les différentes ressources linguistiques utilisées par la méthode que nous proposons dans la section suivante à savoir, les segments conceptuels et les patrons. Ainsi, nous avons dégagé 104 segments conceptuels et 201 patrons avec ou sans marqueurs de rectification. Les résultats de cette étude qualitative nous donne une idée d'une part, sur la complexité et la variété des structures syntaxiques de l'arabe parlé, et d'autre part sur la complexité et la variété des disfluences dans l'arabe parlé.

4.2 Principe de la méthode

Nous proposons de traiter les disfluences au niveau de la compréhension de la parole vu que nous ne disposons pas d'un logiciel *open source* pour la reconnaissance vocale arabe. Ainsi, la méthode proposée est basée sur l'approche à base de patrons et combine la technique de la reconnaissance de patrons à une segmentation conceptuelle des énoncés. L'originalité de cette méthode réside dans l'utilisation des segments conceptuels et la technique de la reconnaissance de patrons pour le traitement des disfluences. Généralement, les travaux qui se basent sur une analyse partielle des énoncés utilisent les *chunks* (purement syntaxiques) qui ont prouvé leur performance dans le traitement de l'écrit. Cependant, pour le traitement de l'oral et pour un domaine limité, comme le notre, nous jugeons que l'utilisation des segments conceptuels (purement sémantiques) est plus intéressante que l'utilisation des *chunks* et peut donner de bons résultats. Par ailleurs, nous sommes convaincus que la segmentation conceptuelle peut aussi jouer un rôle important dans la résolution des disfluences imbriquées.

Par ailleurs, notre méthode consiste en trois principales étapes à savoir, l'étape de découpage des énoncés arabes en segments conceptuels, l'étape de délimitation des segments disfluents et l'étape de détection des disfluences ainsi que leur correction. Afin de bien expliquer notre méthode, nous proposons de prendre à titre d'exemple l'énoncé étiqueté sémantiquement (1). Cet énoncé contient des disfluences qui seront détectées puis corrigées à travers les étapes de la méthode proposée.

⟨$Station$, صفاقس⟩ ⟨$Marq\_Station\_Départ$, من⟩ ⟨$Type\_Billet$, ذهاب⟩ ⟨$Marq\_Billet$, تذكرة⟩ ⟨Mot_Ref_Prix, ثمن⟩ ⟨$Demande$, أراد⟩ **(1)**
⟨$Station$, صفاقس⟩ ⟨$Marq\_Station\_Départ$, من⟩ ⟨$Type\_Billet$, ذهاب – إيّاب⟩ ⟨$Marq\_Billet$, تذكرة⟩ ⟨$Marq\_Rectification$, عفو⟩
⟨$Marq\_Minute$, دقيقة⟩ ⟨$Nombre$, 33⟩ ⟨$Nombre$, 11⟩ ⟨$Marq\_Heure$, على – ساعة⟩ ⟨$Station$, تونس⟩ ⟨$Marq\_Station\_Arrivée$, إلى⟩

- **Découpage en segments conceptuels**

Cette étape consiste à découper l'énoncé en segments conceptuels. Par segment conceptuel, nous entendons la succession des classes de mots dans un seul segment référant à un concept bien déterminé ; c'est-à-dire, la séquence de classes de mots exprimant une même unité de sens (Bousquet-Vernhettes, 2002). Ainsi, une séquence de mots réalisant un segment conceptuel est une instance de ce segment conceptuel. Par exemple, la séquence de mots « من تونس » [*mn twns*] (de Tunis) est une instance du segment conceptuel *Départ* et fait référence au concept *Départ*. Les segments conceptuels se divisent en trois types à savoir, les *illocutoires*, les *référentiels* et les *rebuts* (Bousquet-Vernhettes, 2002). Le type illocutoire fait référence à la théorie des actes du langage (*e.g., Demande*, *Début_Dialogue*, etc.). Le segment « مرحبا » [*mrHbA*] (salut) est le segment conceptuel *Début_Dialogue*. Quant au type référentiel, il permet de représenter le domaine de l'application (*e.g., Heure_Départ*, *Départ*, *Destination*, etc.). Le segment « إلى صفاقس » [*IlY SfAqs*] (à Sfax) est un segment conceptuel *Station_Arrivée*. Le type rebut regroupe les mots et les groupes de mots considérés comme inutiles pour la compréhension de l'énoncé (*e.g., Bruit, Digression*, etc.).

La première phase dans cette étape consiste à repérer les différents marqueurs qui figurent dans un énoncé. Ces marqueurs se définissent comme étant des indicateurs permettant l'identification d'un segment conceptuel. Généralement, les marqueurs ne portent pas d'informations pertinentes mais ils facilitent l'extraction de ces informations. En langue arabe, un marqueur peut être un pré-marqueur, un post-marqueur ou un pré-post marqueur à la fois.

Un pré-marqueur est suivi par l'information pertinente (*i.e.,* le cas sémantique) qu'il marque. Par exemple, le pré-marqueur « على-ساعة » [*ElY-sAEp*] (à-heure) marque le cas sémantique *Nombre_Heure* d'où il sera suivi d'une information sur le nombre d'heures. De la même façon, un post-marqueur est précédé par le cas sémantique qu'il marque. Par exemple, le post-marquer « دقيقة » [*dqyqp*] (minute) est toujours précédé par le cas sémantique *Nombre_Minute*. Un pré-post marqueur est à la fois précédé et suivi par deux cas sémantiques. Par exemple, le pré-post marqueur « تذكرة » [*t*krp*] (billet) est généralement précédé par le cas sémantique *Nombre_Billet* et suivi par le cas sémantique *Type_Billet*. Ainsi, le repérage des marqueurs permet l'identification et la délimitation des segments conceptuels puisque ces derniers sont constitués par les marqueurs et les cas sémantiques qu'ils marquent. Par exemple, le marqueur de station de départ « من » [*mn*] (de) et le cas sémantique *Station* « صفاقس » [*SfAqs*] (Sfax) constituent ensemble le segment conceptuel *Départ* dans l'énoncé (1).

La deuxième phase dans cette étape consiste à raffiner les étiquettes sémantiques des mots composant les segments conceptuels. Ce raffinement est basé essentiellement sur l'étiquette sémantique non raffinée du mot ainsi que le segment conceptuel où se trouve ce mot. En effet, la méthode proposée prend en considération le contexte du mot au sein du segment conceptuel. Ainsi, l'étiquette sémantique non raffinée *Station* peut avoir plusieurs raffinements possibles selon le segment conceptuel où se trouve le mot qui a cette étiquette (*i.e., Station_Départ, Station_Arrivée* ou *Station_Escale*). Par exemple, le mot « صفاقس » [*SfAqs*] (Sfax) du segment conceptuel *Départ* de l'énoncé (1) aura comme étiquette sémantique raffinée *Station_Départ* au lieu de *Station,* ce qui précise exactement le rôle sémantique joué par ce mot (dans ce cas, la station de départ). L'énoncé (2) est le résultat du découpage en segments conceptuels de l'énoncé (1).

$$\left\{\begin{matrix}\text{من صفاقس}\\ [mn\ SfAqs]\\ (de\ Sfax)\end{matrix}\middle|Départ\right\}\quad\left\{\begin{matrix}\text{تذكرة ذهاب}\\ [t*krp\ *hAb]\\ (billet\ aller\ simple)\end{matrix}\middle|Type\_Billet\right\}\quad\left\{\begin{matrix}\text{أراد ثمن}\\ [OrAd\ vmn]\\ (vouloir\ prix)\end{matrix}\middle|Demande\_Prix\right\}$$
$$\left\{\begin{matrix}\text{من صفاقس}\\ [\ mn\ SfAqs]\\ (de\ Sfax\ )\end{matrix}\middle|Départ\right\}\quad\left\{\begin{matrix}\text{تذكرة ذهاب – إياب}\\ [t*krp\ *hAb-Iy{\sim}Ab\ ]\\ (billet\ \ aller-retour)\end{matrix}\middle|Type\_Billet\right\}\quad\left\{\begin{matrix}\text{عفو}\\ [Efw]\\ (pardon)\end{matrix}\middle|Rectification\right\}$$
$$\left\{\begin{matrix}\text{33 دقيقة}\\ [33\ dqyqp]\\ (33\ minutes)\end{matrix}\middle|Minute\_Départ\right\}\quad\left\{\begin{matrix}\text{على – ساعة 11}\\ [ElY-sAEp\ 11]\\ (\grave{a}-heure\ 11)\end{matrix}\middle|Heure\_Départ\right\}\quad\left\{\begin{matrix}\text{إلى تونس}\\ [IlY\ twns]\\ (vers\ Tunis)\end{matrix}\middle|Destination\right\}$$

(**2**)

- **Délimitation des segments disfluents**

L'étape précédente génère, comme résultat, un énoncé qui est soit complètement segmenté soit partiellement segmenté (*i.e.,* il contient des mots isolés n'appartenant à aucun segment conceptuel, comme le cas du mot « عفو » [*Efw*] (pardon) dans l'énoncé (2)). Dans le cas d'une segmentation complète, l'énoncé est soit correct (*i.e.,* ne contient pas de disfluences) soit incorrect (*i.e.,* contient des disfluences). Ces disfluences ne peuvent être que des autocorrections ou des répétitions (les cas d'hésitations ou d'amorces sont exclus). Dans le cas d'une segmentation partielle, la présence des mots isolés est due à la présence d'une ou de plusieurs disfluences. En effet, ces mots ne peuvent être que des marqueurs de rectification, des hésitations, des amorces ou des mots corrigeant un segment. Pour expliquer ce dernier cas, nous prenons à titre d'exemple le segment « من صفاقس » [*mn SfAqs*] (de Sfax) corrigé par le mot « تونس » [*twns*] (Tunis) dans l'énoncé (3).

$$\left\{ \begin{matrix} \text{تونس} \\ [twns] \\ (Tunis) \end{matrix} \middle| Station \right\} \quad \left\{ \begin{matrix} \text{عفو} \\ [Efw] \\ (pardon) \end{matrix} \middle| Rectification \right\} \quad \left\{ \begin{matrix} \text{من صفاقس} \\ [mn\ SfAqs] \\ (de\ Sfax) \end{matrix} \middle| Départ \right\} \tag{3}$$

Ainsi, l'objectif de cette étape est de délimiter les disfluences et de les regrouper dans un segment que nous avons nommé segment *disfluent*. Cette délimitation facilite par la suite le repérage et la correction des disfluences dans l'étape en aval. Le segment (4) représente le segment disfluent de l'énoncé (3).

$$\left\{ \begin{matrix} \text{من صفاقس عفو تونس} \\ [mn\ SfAq\ Efw\ twns] \\ (de\ Sfax\ \ pardon\ Tunis) \end{matrix} \middle| Disfluent \right\} \tag{4}$$

Pour délimiter le segment disfluent, notre méthode se base sur des règles de délimitation ainsi que sur la présence ou non de mots isolés. Les règles de délimitation prennent en considération les mots au sein des segments, les étiquettes sémantiques des mots, ainsi que les types des segments conceptuels. Dans le cas de présence des mots isolés, les frontières du segment disfluent sont faciles à repérer. En effet, le segment disfluent est composé du mot isolé et des segments qui l'entourent. Le segment (5) représente le segment disfluent dans l'énoncé (2). Ce segment est composé d'une répétition du mot « تذكرة » [*t*krp*] (billet) et du segment « من صفاقس » [*mn SfAqs*] (de Sfax), et d'une autocorrection du type de billet « ذهاب » [**hAb*] (aller simple) par « ذهاب-إياب » [**hAb-Iy~Ab*] (aller-retour).

$$\left\{ \begin{matrix} \text{تذكرة ذهاب من صفاقس عفوتذكرة ذهاب – إياب من صفاقس} \\ [t*krp*hAb\ mn\ SfAqs\ Efw\ *t*krp*hAb-Iy\sim Ab\ mn\ SfAqs] \\ (billet\ aller\ simple\ de\ Sfax\ pardon\ \ billet\ aller-retour\ de\ Sfax) \end{matrix} \middle| Disfluent \right\} \tag{5}$$

Dans le cas de non présence de mots isolés qui ciblent l'analyse, les règles de délimitation sont appliquées sur tout l'énoncé afin de vérifier l'existence ou non de segments disfluents. Ces segments disfluents, s'ils existent, sont constitués par des autocorrections (segments corrigés par d'autres segments) et/ou par des répétitions. L'énoncé (6) représente l'autocorrection du segment « أراد ثمن » [*OrAd vmn*] (vouloir prix) par le segment « أراد وقت قطار » [*OrAd wqt qTAr*] (vouloir horaire train). Cette autocorrection est faite sans la présence d'un marqueur de rectification, d'une amorce ou même d'une hésitation.

$$\left\{ \begin{matrix} \text{أراد وقت قطار} \\ [\ OrAd\ wqt\ qTAr] \\ (vouloir\ horaire\ train\ ) \end{matrix} \middle| Demande\_Horaire \right\} \left\{ \begin{matrix} \text{أراد ثمن} \\ [OrAd\ vmn] \\ (vouloir\ prix) \end{matrix} \middle| Demande\_Prix \right\} \tag{6}$$

Le segment (7) est le segment disfluent de l'énoncé (6) après l'utilisation des règles de délimitation.

$$\left\{ \begin{matrix} \text{أراد ثمن أراد وقت قطار} \\ [OrAd\ vmn\ OrAd\ wqt\ qTAr] \\ (vouloir\ prix\ vouloir\ horaire\ train) \end{matrix} \middle| Disfluent \right\} \tag{7}$$

Le résultat de cette étape est un énoncé découpé à la fois en segments conceptuels et en segments disfluents (si ces derniers existent). Les segments disfluents seront traités au niveau de l'étape suivante.

- **Détection et correction des disfluences**

L'objectif de cette étape est de détecter les disfluences existantes dans les segments disfluents déjà repérés dans l'étape précédente, ensuite, de les corriger. Pour se faire, la reconnaissance de patrons est appliquée. Cette technique consiste à parcourir l'ensemble des patrons afin de déterminer, pour chaque segment disfluent, le patron qui lui correspond. À ce stade d'analyse, nous tenons à signaler que la reconnaissance de patrons se fait d'une manière robuste. En effet, lors de la reconnaissance, un patron peut être retenu même si le segment disfluent comporte un ou plusieurs mots considérés inutiles et qui n'ont pas de correspondants dans le patron. Cette robustesse permet de réduire le nombre de patrons utilisés et de surmonter un problème spécifique à la langue arabe à savoir, la grande taille des énoncés.

Les patrons utilisés concernent les cas de répétitions, d'autocorrections, d'hésitations, d'amorces ou de combinaison des quatre types de disfluences. Ces patrons permettent l'identification ***i)*** de suites de mots répétés (de manière identique : **M**), repris (mots différents jouant le même rôle sémantique : **R**) ou ajoutés (mots neutres : **X**), ***ii)*** d'un terme d'édition (marqueur de reprise : **ET**) et ***iii)*** d'un point d'interruption noté par une barre verticale (|). Ainsi, le patron correspondant au segment disfluent (5) est le suivant :

$$\begin{matrix} \text{صفاقس من ذهاب – إيّاب من صفاقس} & \text{تذكرة ذهاب من صفاقس} \\ \text{M3 M2 R1 M1} & \text{ET} \mid \text{M3 M2 R1 M1} \end{matrix} \quad (8)$$

Le segment disfluent est alors décrit comme la succession d'un *reparandum* (partie du segment qui sera corrigée par la suite), d'un *interregnum* optionnel (marqueur de reprise) et d'un *repair* (partie qui corrige le *reparandum*) comme illustré par l'exemple (9).

$$\begin{bmatrix} \text{تذكرة ذهاب – إياب من صفاقس} \\ [t * krp * hAb - Iy{\sim}Ab\ mn\ SfAqs] \\ (billet\ aller - retour\ de\ Sfax) \end{bmatrix}_{Repair} \quad \begin{bmatrix} \text{عفو} \\ [Efw] \\ (pardon) \end{bmatrix}_{Interregnum} \quad \begin{bmatrix} \text{تذكرة ذهاب من صفاقس} \\ [t * krp * hAb\ mn\ SfAqs\ ] \\ (billet\ aller\ simple\ de\ Sfax) \end{bmatrix}_{Reparandum} \quad (9)$$

Pour la correction proprement dite, le *repair* est gardé ; cependant l'*interregnum* ainsi que le *reparandum* sont supprimés. Le segment résultat subit une analyse similaire à celle de l'étape de découpage en segments conceptuels afin de le segmenter en segments conceptuels et de raffiner les étiquettes sémantiques des mots qui les composent. Les deux segments conceptuels (10) représentent le segment disfluent (5) après correction et découpage en segments conceptuels.

$$\left\{ \begin{matrix} \text{من صفاقس} \\ [mn\ SfAqs] \\ (de\ Sfax) \end{matrix} \middle| Départ \right\} \quad \left\{ \begin{matrix} \text{تذكرة ذهاب – إيّاب} \\ [t * krp * hAb - Iy{\sim}Ab] \\ (billet\ aller - retour) \end{matrix} \middle| Type\_Billet \right\} \quad (10)$$

5 Mise en œuvre et évaluation de la méthode proposée

Nous avons implémenté la méthode proposée avec le langage JAVA sous l'environnement JBuilder 2007. Les patrons et les segments conceptuels sont regroupés dans des fichiers XML. Pour valider notre proposition, nous avons intégré notre système de traitement des disfluences dans le module de compréhension du système SARF (Serveur vocal Arabe des Renseignements sur le transport Ferroviaire). Nous signalons qu'avant l'intégration de notre système au sein du module de compréhension, ce dernier ne traitait pas les disfluences. Par ailleurs, lors de l'évaluation du système SARF, ce dernier a obtenu 71.79% comme taux de *F-Measure* et 18.54% comme taux d'erreurs (Hadrich Belguith et al., 2009).

Pour montrer l'efficacité de notre système de traitement des disfluences et son apport vis-à-vis du système SARF, nous avons pris le même corpus utilisé lors de sa dernière évaluation. Ce corpus a été construit selon la technique du *Magicien d'Oz*. Il est constitué de 2535 énoncés (soit 32520 mots) prononcés d'une manière spontanée. Les occurrences d'apparition des différents types de disfluences dans ce corpus d'évaluation sont : 859 autocorrections, 738 répétitions, 388 hésitations et 342 amorces.

La nouvelle évaluation du module de compréhension de SARF a montré que les taux de rappel, de précision et de *F-Measure* sont respectivement de 79.23%, 74.09% et 76.57%. Le temps moyen d'exécution d'un énoncé, de 12 mots, est d'environ 0.394 secondes. Le taux d'erreurs est de 12.63% ; ce qui représente une diminution de 5.91% d'erreurs. À la lumière de ces résultats encourageants, nous jugeons que la méthode proposée est efficace pour le traitement des disfluences dans l'oral arabe même s'il reste à étudier les cas d'échec en vue d'améliorer les résultats obtenus. Ces cas d'échec (12.63%) peuvent être résumés comme suit :

- Le premier cas d'échec est causé par la présence de mots hors-vocabulaire dans les énoncés traités. Ce phénomène est dû à la performance du système de reconnaissance utilisé et non pas au système de traitement de disfluences que nous avons réalisé. Prenons à titre d'exemple le mot « سقف » [*sqf*] (toit) dans l'énoncé (11). Ce mot représente un mot hors-vocabulaire et aura « HV » comme étiquette sémantique puisqu'il n'appartient pas à notre domaine applicatif. Les mots hors-vocabulaire sont difficiles à comprendre par le module de compréhension, ce qui entrave l'étiquetage sémantique et complique la tâche de segmentation conceptuelle. La présence de ce type de mots lors de la segmentation entraîne une perturbation dans la sélection du bon segment conceptuel ; ce qui fausse la délimitation des segments disfluents et par conséquent rend la détection des types de disfluences impossible.

$$\begin{bmatrix} \text{من صفاقس عفو من سقف} \\ [mn\ SfAqs\ Efw\ mn\ sqf] \\ (\ de\ Sfax\ pardon\ de\ toit\ ) \end{bmatrix} \quad (11)$$

- Le deuxième cas d'échec est causé par des erreurs au niveau étiquetage sémantique des énoncés et non pas par le système de traitement des disfluences. En effet, une mauvaise assignation des étiquètes sémantiques, aux mots d'un énoncé, peut entrainer un mauvais découpage de ce dernier et par suite provoquer des erreurs de détection des disfluences. Ce cas d'ambiguïté est dû principalement à la non distinction entre les marqueurs de négation et les marqueurs de rectification. À titre d'exemple, le mot « لا » [*lA*] (non) de l'énoncé (12) peut jouer deux rôles sémantiques à savoir, un marqueur de négation ou un marqueur de rectification. Ce type d'ambiguïté est résolu au niveau raffinement des étiquettes sémantiques lors de la première étape du traitement des disfluences. Cependant, à l'état actuel, notre système est incapable d'affecter au mot « لا » [*lA*] (non) la bonne étiquette sémantique. En effet, le système de traitement des disfluences doit avoir des informations sur le contexte de l'énoncé (12) au sein du dialogue. Chose qui n'est pas encore prise en considération dans la méthode proposée.

$$\begin{bmatrix} \text{من صفاقس لا من تونس} \\ [mn\ SfAqs\ lA\ mn\ twns] \\ (\ de\ Sfax\ \ non\ de\ Tunis\ ) \end{bmatrix} \quad (12)$$

- Le troisième cas d'échec est dû à la présence d'énumérations dans les énoncés. En effet, l'aspect structurel d'une énumération est très proche de celui d'une autocorrection. À titre d'exemple, l'énoncé (13) présente une énumération de deux types de billet à savoir, « ذهاب » [**hAb*] (aller simple) et « ذهاب-اياب » [**hAb-AyAb*] (aller retour) que notre système a considéré comme étant un cas d'autocorrection. Ce phénomène fausse les résultats et reste non encore résolu dans la plupart des systèmes de traitement des disfluences.

$$\begin{bmatrix} \text{ماهو سعر تذكرتين ذهاب ذهاب – اياب} \\ [mAhw\ sEr\ t * krtyn\ * hAb * hAb - AyAb] \\ (\ Quel\ est\ le\ prix\ de\ deux\ tickets\ aller\ simple\ \ aller\ retour\ ) \end{bmatrix} \quad (13)$$

6 Conclusion et perspectives

La compréhension de la parole spontanée est un thème de recherche très riche même s'il reste encore de nombreux progrès à faire. En particulier, la parole arabe qui n'a pas fait l'objet de plusieurs travaux de

recherche (comparée à d'autres langues comme l'anglais et le français). Cela est dû d'une part, au manque d'outils nécessaires (*i.e.,* principalement les corpus oraux) et d'autre part, aux spécificités de l'oral arabe (*i.e.,* diversité dialectale de l'arabe parlé, irrégularité de l'ordre des mots, etc.).

Dans cet article, nous avons proposé une méthode originale pour le traitement des disfluences dans le cadre de la compréhension de la parole arabe spontanée. Cette méthode repose sur une approche à base de patrons et combine la technique de la reconnaissance de patrons à une segmentation conceptuelle. Ainsi, un énoncé étiqueté sémantiquement passe par trois étapes de traitement : étape du découpage en segments conceptuels, étape de délimitation des segments disfluents et étape de détection et de correction des disfluences. Les résultats que nous avons obtenus sont encourageants (*F-Measure* égale à 76.57%).

Comme perspectives, nous envisageons d'étudier les phénomènes de négation et d'énumération en vue d'apporter des solutions quant à leur détection. Cela évitera de les confondre avec les disfluences. Aussi, nous projetons de traiter les mots hors-vocabulaire (*i.e.,* les mots inconnus par le système de compréhension et les mots mal-reconnus par le système de la reconnaissance vocale).

Références

Bahou Y., Hadrich Belguith L., Ben Hamadou A. (2008). Towards a Human-Machine Spoken Dialogue in Arabic. *LREC'08, Workshop HLT within the Arabic World: Arabic Language and local languages processing Status Updates and Prospects*, Marrakech, Morocco.

Bear J., Dowding J., Shriberg E. (1992). Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog, *In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Deleware, USA.

Bouraoui J.-L. (2009). Traitement automatique de dysfluences dans un corpus linguistiquement constraint, *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles*, *TALN'09,* Senlis, France.

Bousquet-Vernhettes C. (2002). Compréhension Robuste de la Parole Spontanée dans le Dialogue Oral Homme-Machine – Décodage Conceptuel Stochastique, *Thèse de doctorat à l'Université de Toulouse III–Paul SABATIER*, France.

Bove R. (2008). A Tagged Corpus-Based Study for Repeats and Self-Repairs Detection in French Transcribed Speech, *Proceedings of the 11th International Conference on Text, Speech and Dialogue, TSD'08,* Brno, Czech Republic.

Core M., Schubert L. (1999). A Model of Speech Repairs and Other Disruptions, *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems,* Cape Cod, MA, USA.

Hadrich Belguith L., Bahou Y., Ben Hamadou A. (2009). Une méthode guidée par la sémantique pour la compréhension automatique des énoncés oraux arabes. *International journal of information sciences for decision making, ISDM'09,* volume 35.

Heeman P.A., Allen J.F. (1996). Combining the Detection and Correction of Speech Repairs, *In Proceedings of International Conference of Spoken Language Processing,* Philadelphia, PA, USA.

Kurdi Z. (2003). Contribution à l'analyse du langage oral spontané, *thèse de doctorat à l'université Joseph Fourier*, France.

Mckelvie D. (1998). The syntax of disfluency in spontaneous spoken language. *Technical Report HCRC/RP-95*, Edinburgh University, Edinburgh, Scotland.