TAGGAR : Un analyseur morphosyntaxique destiné à la synthèse vocale de textes arabes voyellés

Z. Zemirli, S. Khabet

Institut National d'Informatique
BP 68M – 16309 Oued-smar Alger – ALGERIE
z_zemirli@ini.dz, khabet_s@hotmail.com

Résumé – Abstract

La qualité d'un système de synthèse vocale à partir du texte dépend du naturel, de l'intelligibilité de la parole générée et des caractéristiques propres à la voix produite. Ces caractéristiques dépendent des techniques et des méthodes de synthèse, mais également du soin apporté à la ***modélisation linguistique*** et prosodique. Plusieurs travaux soulignent le fait que des ***structures linguistiques*** entretiennent des liens étroits avec les ***réalisations prosodiques*** (Mertens et al, 2001). Concernant la langue arabe, les modèles se basent sur la structure syllabique des mots, l'accentuation, la notion de marqueurs intonatifs manuels (interrogatif principalement) et très peu sur des informations syntaxiques. Notre objectif est de développer un système de marquage syntaxique automatique simple mais pouvant améliorer le contour intonatif des phrases synthétisées. L'analyseur morphosyntaxique TAGGAR réalise un étiquetage grammatical des mots afin d'assurer l'accentuation des mots, le groupage syntaxique, l'insertion des groupes de souffle et le calcul intonatif des phrases synthétisées.

The quality of a text to speech synthesis depends on the naturalness, on the intelligibility of the speech generated and the specific characteristics to the produced voice. These characteristics depend on the techniques and the methods of synthesis, but also on the care taken to ***linguistic*** and prosodic modeling. Several works underlines the fact that ***linguistic structures*** maintain the close links with the ***prosodic achievements***. Concerning the Arab language, the models are based on the syllabic structure of the words, the stressing, the concept of markers intonates (interrogative mainly) and very little on syntactic information. Our objective is to develop a simple system of automatic syntactic marking but being able to improve intonate contour of the synthesized sentences. Analyzer morphosyntactic TAGGAR carries out a grammatical labeling of the words in order to ensure the stressing of the words, the syntactic grouping, the insertion of the groups of breath and the intonate calculation of the synthesized sentences.

Keywords – Mots Clés

Synthèse vocale à partir de textes arabes voyellés, analyse morphosyntaxique, prosodie.
Arabic Text To Speech Synthesis, arabic diacritized texts, morphosyntactic analysis, prosody

1 Introduction

TAGGAR est un analyseur morphosyntaxique spécialement développé pour la synthèse vocale arabe à partir de textes arabes voyellés. Il a pour objectif de produire une ***analyse syntaxique spécifique*** en vue de son utilisation en ***temps réel*** dans un système de ***synthèse vocale***. L'analyse syntaxique au sens classique et complet n'est pas une fin en soi, mais elle a pour objectif de contribuer à améliorer la qualité de la synthèse vocale en la rendant aussi acceptable que possible. Il faut que l'analyseur soit adapté aux contraintes inhérentes au fonctionnement d'un système de synthèse vocale, en terme de performances mais également en terme de résultats (rapide, robuste et déterministe). Cet outil est important dans un système de synthèse de la parole à partir du texte, car l'insertion des pauses (obtenues après la détermination des groupes de souffle) et la génération des marqueurs prosodiques, ne peut être faite que si l'on dispose d'un minimum d'information grammaticale sur chaque mot de la phrase. Les textes soumis à l'entrée de notre système sont supposés correctement voyellés. L'étiqueteur a pour rôle d'identifier la catégorie d'un mot donné et de lui affecter une étiquette morphosyntaxique (verbe inaccompli, nom accusatif, etc.). Le problème majeur qui se pose est celui des éléments affixés qui sont liés au mot (suffixes et/ou préfixes). Le système procède en premier lieu à un découpage éventuel du mot. Le principe du découpage est d'isoler les préfixes et les suffixes du mot; La partie restante correspond à la base. Tout système d'analyse morphosyntaxique se compose de deux parties essentielles à savoir, une base de données lexicales qui contient et décrit les ***ressources lexicales*** et un ***analyseur***. TAGGAR a été implémenté dans le système ARAVOICE (Zemirli et al 2003) de synthèse vocale arabe à partir de textes arabes voyellés qui utilise le moteur de synthèse et la notation des phonèmes Mbrola. (http://tcts.fpms.ac.be/synthesis/).

2 Ressources lexicales

Les besoins en ressources lexicales nécessaires au développement d'un système de synthèse vocale de bonne qualité ne sont plus à démontrer. Les ressources électroniques en langue arabe ne sont pas encore disponibles et diffusées en vue d'un traitement automatique bien que ne nombreux travaux de laboratoire aient été réalisé (Dichy et al, 2002). Nous avons défini pour notre travail des ressources, certes partielles, mais que nous jugeons suffisantes pour atteindre nos objectifs. Les ressources lexicales sont composées de quatre lexiques qui fournissent toutes informations grammaticales nécessaires à l'analyseur. Le lexique des mots outils recouvre l'ensemble des particules (les prépositions, les conjonctions de coordination, les particules d'interrogation, etc.). Le lexique des mots spécifiques contient l'ensemble des mots non dérivables. Le lexique des schèmes générateurs est relatif uniquement aux verbes. Enfin, le dernier lexique contient l'ensemble des éléments affixés. ***Les antéfixes et les préfixes*** sont des morphèmes qui entrent dans la formation des mots. Ils s'ajoutent aux noms et aux verbes et se placent devant la base. Afin de faciliter la phase de segmentation des verbes et des noms, nous les avons prétraités suivant leur taille, le temps et les combinaisons attestées. Les tailles varient entre 1 et 2 lettres en fonction du temps du verbe (وَسَمِعَ /wasamiHa/, فَسَيَسمَعُ /fasajasmaHu/, etc.). Les combinaisons des antéfixes et des préfixes sont ordonnées selon leur longueur que l'on peut trouver devant un nom. Leur taille varie entre 1 à 5 lettres (المِحرَاثُ /almiXraaTu/, بِالمِحرَاثِ /bilmiXraaTi/, etc.). ***Les suffixes et les postfixes*** peuvent se raccorder aux verbes comme aux noms après la base. Des modules de conjugaison permettent de générer les schèmes verbaux avec les suffixes correspondants. Cinq modules de

conjugaison sont utilisés pour l'accompli actif et passif, l'inaccompli actif et passif et l'impératif. (Par exemple avec *mode*=accompli, *schème*=1 (**فَعَلَ**) et *personne*=12 (**هُم**), le *résultat* est : **فَعَلُوا**). Les suffixes nominaux sont insérés dans une table. Ils s'utilisent aussi bien avec les noms générés à partir de racines dérivées mais, également avec des mots outils et des mots spécifiques. Les postfixes qui terminent un verbe conjugué sont détectés en relation avec le temps du verbe. ***Les mots outils*** sont regroupés selon leur fonction dans des tables (table des particules de conjonction, des pronoms personnels, etc.). Ils forment un ensemble fini et ces tables contiennent toutes les formes associées : simples **فِي** /fii/, suffixées **فِيهِ** /fiihi/, **فِيهَا** /fiihaa/, affixées **وَفِي** /wafii/, ou les eux à la fois **وَفِيهِم** /wafiihim/). ***Les mots spécifiques*** sont des noms qui n'ont pas de racine dans la langue arabe. Ce sont des mots non dérivables et certains sont inflexibles. Il est alors nécessaire de les recenser afin de les traiter correctement. Parmi les mots inflexibles, on trouve les noms propres (noms de pays ou de personnes) ou des noms communs. On peut citer **فاطِمَةُ** /faatimatu/, **صَيدَحُ** /s.a.jdaXu/, **مِصرُ** /mis.r/, **مَكَّةُ** /makkata/, **عِيسَى** /Hiisaa/, ou les noms commençant par **عَبدُ** /Habdu/ suivi du nom ou d'une épithète d'Allah comme **عَبدُ اللَّهِ** /Habdu ?allahi/. ***Le lexique des schèmes verbaux*** est subdivisé en trois sous lexiques correspondant aux trois temps du verbe. Trois modules de conjugaison sont utilisés pour générer les formes conjuguées des schèmes des trois sous lexiques. Nous avons défini 72 schèmes pour l'accompli (53 schèmes pour la voix active et 19 schèmes pour la voix passive), 55 schèmes pour l'inaccompli (36 schèmes pour la voix active et 19 schèmes pour la voix passive), et 44 schèmes pour l'impératif.

3 Définition du jeu d'étiquettes de TAGGAR

Notre objectif est d'effectuer les traitements suivants : étiquetage des mots, ***regroupement syntaxique*** à partir des mots étiquetés, regrouper les groupes syntaxiques en groupes de souffle (insertion automatique des pauses) et disposer d'informations sur des mots pouvant servir à la production de contours prosodiques particuliers (exclamation, interrogation, etc.). Du point de vue ***prosodique*** le ***groupe syntaxique*** que nous allons définir ne ***peut être découpé*** et ne comportera ***pas de pause***. Nos textes sont entièrement voyellés et notre premier objectif est de regrouper les mots en groupes syntaxiques. Les groupes syntaxiques auront comme élément central un ***verbe*** ou un ***nom*** qui seront reliés par des ***particules*** de coordination, de relation, du génitif, des pronoms, etc. Ceci permet de dégager trois grandes classes d'étiquettes : les particules, les verbes et les noms. Un verbe sera caractérisé par son temps (accompli/inaccompli et impératif) et son mode (actif ou passif) qui renferment des informations utiles à la génération de la prosodie. Les flexions casuelles des noms : sujet, accusatif et génitif ainsi que leur type déterminés/indéterminés jouent un rôle déterminant dans la classification des étiquettes de cette classe. Partant de ces réflexions, nous avons défini pour nos besoins 35 étiquettes grammaticales. Elles se répartissent en trois grandes catégories : 4 étiquettes pour les particules, 16 étiquettes pour les verbes, et 15 étiquettes pour les noms. Baloul (Baloul et al, 2002) définit 21 étiquettes, il considère la conjonction de coordination **'و'** comme étant une entité lexicale séparée du mot qui la suit! Il n'utilise pas les différents temps du verbe ni les formes passives et actives. Or, ces informations sont pertinentes pour le calcul des contours intonatifs des phrases synthétisées. Quand à (Débili et al, 2002), (El-Kareh, Al-Ansary, 2000), (Freeman, 2001), (Khoja et al, 2001), (Ouersighni, 2001), ***leurs importants travaux*** dans le traitement automatique de la langue arabe utilisent un ***grand nombre d'étiquettes grammaticales*** (606 pour Débili) qui ***n'ont jamais été utilisées dans un système de synthèse vocale arabe complet fonctionnant en temps réel.***

La particule accompagne un nom ou un verbe. Notée (P), elle se trouve au début d'un groupe syntaxique (هَذِهِ الْمُذِيعَةُ). Notée (PR), elle débute un groupe syntaxique et à un effet liant avec le groupe syntaxique précédent (الَّتِي تُقَدِّمُ). Certaines particules (PF) ne peuvent pas débuter un groupe syntaxique (فَقَطَ). Enfin, certaines notées (PS) peuvent être suffixées (مَعَكُم) Le jeu d'étiquettes des verbes comporte trois sous-classes représentant les trois temps du verbe. Exemple : VerbeAcc, VerbeAccSuff, VerbeAccPas, etc. L'indication de la préfixation est utilisée dans l'assignation de l'accent local des mots. Les verbes à l'impératif comportent un contour intonatif permettant de mettre en relief l'indication d'ordre par une forte accentuation. L'indication de la voix passive ou active est utilisée dans la détermination de la hauteur de l'accentuation locale des verbes (Zemirli et al, 2003). Le jeu d'étiquettes des noms est tiré des trois cas de flexions des noms sujet, accusatif et génitif. Dans les trois cas, un nom peut être déterminé ou indéterminé, préfixé et/ou suffixé. Exemple : NomDetSujet, NomDetSujetPref, NomIndSujet, etc. Ces informations sont utilisées par le système ARAVOICE dans la phase de regroupement syntaxique.

4 Analyseur

L'ordre de traitement des mots d'un texte est important car, il permet de minimiser les erreurs d'étiquetage. Le traitement se fait dans l'ordre suivant : Analyse des mots outils et des mots spécifiques, analyse des formes verbales et enfin, analyse des formes nominales. L'ensemble des mots outils regroupe les particules qui sont bien identifiées et dont le nombre est bien déterminé. Leurs formes sont pratiquement fixes, c'est-à-dire qu'ils gardent toujours la même forme sauf pour quelques uns qui peuvent être affixés par un suffixe et un ou plusieurs préfixes. Exemple : مَعَكَ (مَعَ + كَ). Le traitement des mots spécifiques est identique à celui des mots outils. La différence réside après les étapes de détection des éléments affixés. Nous utilisons un module spécifique aux noms qui a pour rôle de détecter et de supprimer la déclinaison d'une forme nominale. Le processus d'analyse des formes dérivées verbales est le suivant : Extraction du ou des préfixes, extraction du ou des suffixes avec vérification de la compatibilité entre les différents affixes puis, projection de la base restante sur les schèmes générateurs des formes verbales dérivées. La dernière phase et l'analyse des formes nominales. Les flexions casuelles des cas sujet, accusatif et génitif sont les principaux signes qui guident l'attribution des étiquettes aux noms. On recherche selon dans l'ordre suivant, les noms déterminés simples ou préfixés, les noms indéterminés préfixés ou non puis les noms déterminés par annexion ou suffixation d'un pronom personnel. Lorsque l'analyse échoue, on attribue l'étiquette temporaire (NomDet) si le nom comporte un article sinon, une étiquette temporaire *vide* est attribuée. L'attribution de l'étiquette finale se fera après une analyse contextuelle des étiquettes voisines. L'étiquetage de textes arabes même entièrement voyellés peut aboutir à des cas d'ambiguïtés d'étiquetage. Par exemple, أحمدُ est un verbe à l'inaccompli ou un nom au cas sujet. أَنَّ est une particule de Nasb ou un verbe à l'accompli (il se plaint). Des règles contextuelles de désambiguïsation sont utilisées pour assigner une étiquette en fonction des contextes en agissant sur les mots repérés. Par exemple, أَنَّ sera étiqueté HarfNasb dans : تَعلَمُونَ أَنَّ السَمَاءَ لاَ تُمطِرُ ذَهَبًا وَلاَ فِضَّةً. et il recevra l'étiquette VerbeAcc dans : أَنَّ المَرِيضُ مِنَ الأَلَمِ الشَّدِيدِ.

5 Applications de TAGGAR

Nous allons dans ce qui suit décrire de manière succincte quelques applications directes de l'analyseur TAGGAR telles que le groupage syntaxique, l'insertion des groupes de souffle et la détermination de l'accentuation locale des mots ou la génération du contour intonatif.

L'objectif étant de montrer l'apport de TAGGAR dans un système de synthèse vocale. TAGGAR est utilisé pour l'étiquetage, le groupage syntaxique et l'insertion des groupes de souffle dans (Zemirli et al 2003). L'étiquetage de la phrase : شَرِبَ طِفلٌ لَطِيفٌ الحلِيبَ اللَّذِيذَ وَأَكَلَ رَجُلٌ . البُرتُقَالَةَ المُنعِشَةَ fournit le résultat suivant : شَرِبَ (VerbeAcc) طِفلٌ (NomIndSujet) لَطِيفٌ (NomIndSujet) الحلِيبَ (NomDetAccusatif) اللَّذِيذَ (NomDetAccusatif) وَأَكَلَ (VerbeAccPref) رَجُلٌ (NomIndSujet) البُرتُقَالَةَ (NomDetAccusatif) المُنعِشَةَ (NomDetAccusatif). (Ponct). A partir des quatre groupes syntaxiques défini (GV : groupe verbal, GNS : groupe sujet, GNA : groupe accusatif et GNG : groupe génitif), nous effectuons le groupage syntaxique suivant : (شَرِبَ) (طِفلٌ لَطِيفٌ) (الحلِيبَ اللَّذِيذَ) (وَأَكَلَ) (رَجُلٌ) (البُرتُقَالَةَ المُنعِشَةَ.). Puis, nous procédons à l'insertion des pauses (#) entre les groupes de souffle à l'aide des règles définies. (شَرِبَ) (طِفلٌ لَطِيفٌ) (الحلِيبَ اللَّذِيذَ) # (وَأَكَلَ) (رَجُلٌ) (البُرتُقَالَةَ المُنعِشَةَ.). Les groupes syntaxiques défini permettent de définir des remontées intonatives. Les étiquettes produites sont utilisées pour définir aussi bien la syllabe accentuée des mots mais aussi la hauteur de cette accentuation.

6 Evaluation

A partir d'un corpus de 5563 mots, TAGGAR a engendré un taux d'erreur de 2% sur les étiquettes qui a entraîné seulement 1% d'erreurs sur les frontières de groupes syntaxiques. Près de 98% des pauses insérées automatiquement sont correctement placées.

Mots correctement étiquetés	5451	97,99
Etiquettes incorrectes	112	02,01
Groupes syntaxiques	2547	
Frontières syntaxiques correctes	2521	98,98
Frontières syntaxiques incorrectes	26	01,02
Pauses insérées	337	
Pauses correctement insérées	330	97,92
Pauses incorrectement insérées	7	02,08

Tableau 1. Evaluation de TAGGAR

Le mot بطَاقَاتٍ dans la phrase بطَاقَاتٍ كَبيرَةٍ. peut-être interprété comme le mot بطَاقَاتٍ ou bien طَاقَاتٍ + بِ. Cela n'a aucune incidence sur le regroupement, on obtient le même groupe nominal génitif. (بطَاقَاتٍ كَبيرَةٍ). Le processus d'étiquetage morphosyntaxique repose sur la priorité de l'étiquetage des verbes sur les noms. Certains noms ont une structure qui s'apparente à celle des schèmes des verbes que nous avons définis. Le mot أَكْثَرُ est étiqueté comme un verbe alors qu'il correspond à un nom sujet indéterminé. Mais, cette erreur n'a aucune incidence sur le regroupement syntaxique ni sur l'insertion de la pause après un groupe de souffle. La phrase وَإِنْ اجْتَمَعَ شَخْصَانِ أَوْ أَكثَرُ مِنْ أَهْلِ الْحَضَانَةِ. est regroupée en (وَإن)(اجتَمَعَ)(شَخصَان)(أَو أَكثَرُ)#(مِن أَهل الحَضَانَةِ.). La langue arabe comporte un certain nombre de mots inflexibles, leur flexion casuelle ne subit aucun changement par l'influence d'un régissant. Par exemple le nom propre de femme حَذَام est inflexible dans les trois cas. Dans جَاءَت حَذَام, حَذَام est un nom au cas sujet. Dans مَرَرتُ بِحَذَام, حَذَام et au cas génitif et dans رَأَيتُ مِن بَينِ الأَشجَارِ حَذَام, حَذَام au cas accusatif. Certaines omissions de la voyelle finale dans des mots ont engendré quelques erreurs d'étiquetage. Certaines erreurs ont été rectifiées à l'aide des constituants du mot et des règles contextuelles. Le mot بالذكر bien qu'étant amputé de sa voyelle finale est correctement étiqueté comme NomDetGenitPref.

7 Conclusion

L'analyseur morphosyntaxique TAGGAR réalise un étiquetage des mots d'un texte pour répondre aux objectifs suivants : Regrouper les mots en groupes syntaxiques, insérer des pauses entres des groupes de souffle et utiliser les étiquettes pour le calcul du contour intonatif des mots et des phrases dans le système ARAVOICE de synthèse vocale à partir de textes arabes voyellés. 2% d'erreurs d'étiquetage ont engendré 1% de mauvais regroupements syntaxiques. Certaines erreurs d'étiquetage n'ont pas eu d'influence sur le découpage en groupes syntaxiques. Les erreurs les plus fréquentes que nous avons recensé sont celles provoquées par les noms qui sont étiquetés comme des verbes et les noms propres inflexibles. Le système d'étiquetage développé utilise des ressources linguistiques partielles qui pourraient être complétées par un lexique de racines verbales et un dictionnaire des noms propres. Néanmoins, ce coûteux investissement en nouvelles ressources ne semble pas justifié en terme d'amélioration de gain de performance.

Références

Baloul S., Alissali M., Baudry M., Boula de MAreüil P., Interface syntaxe-prosodie dans un système de la parole à partir du texte en arabe, *XXIVèmes JEP*, Nancy, 329-332.

Débili F., Achour H., Souici E. (2002), La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique, *Correspondances de l'IRMC*, N° 71, 10-28.

Dichy J., Braham A., Ghazali s., Hassoum M. (2002), La base de connaissances linguistiques DIINAR.1, *International Symposium on The Processing of Arabic*, Tunis, 45-56.

Dutoit T., Pagel V., Pierret N., Bataille F., Van Der Vrecken O. (1996), The MBROLA project : towards a set of high quality speech synthetizers free of use for non commercial purpose, *icslp'96*.

El-Kareh S., Al-Ansary S. (2000), An Arabic Interactive Multi-feature POS Tagger. In *Proceedings of the, ACIDCA conference*, Monastir, Tunisia, 204-210.

Freeman A., (2001), Brill's POS tagger and a Morphology parser for Arabic, *Ph.D., Department of Near Eastern Studies*, Michigan, USA

Khoja S., Garside R., Knowles G. (2001), A Tagset for the Morphosyntactic Tagging of Arabic, *Proceedings of the Corpus Linguistics 2001*, Lancaster University (UK), Volume 13 - Special issue, 341.

Mertens P., Goldman J.P., Wehrli E., Gaudinat A. (2001), la synthèse de l'intonation à partir de structures syntaxiques riches", *TAL*, 42/1.

Ouersighni R. (2001), A major offshoot of the DIINAR-MBC project: *AraParse*, a morphosyntactic analyzer for unvowelled Arabic texts, *in ACL 39th Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect,* Toulouse, 9-16.

Zemirli Z., Obrecht R.A., Henni A., Sellami M. (2003), Aravoice: an Arabic Text-To-Speech System, *in proceedings of SPECOM'03*, Moscow, 170-177.