Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue Amazighe

Fadoua Ataa Allah, Hassan Jaa

CEISIC, IRCAM
{ataaallah, jaa}@ircam.ma

Résumé – Abstract

L'utilisation des corpus en traitement automatique des langues, notamment ceux annotés morphosyntaxiquement, est devenue une étape indispensable pour l'élaboration des outils linguistiques contribuant à l'informatisation des langues.

Dans l'objectif de promouvoir et d'informatiser la langue amazighe, nous sommes amenés à construire un corpus annoté morphosyntaxiquement. Ce qui a suscité le développement d'un outil fournissant de l'aide et de l'assistance aux linguistes.

The use of corpora in natural language processing, especially those annotated morphosyntactically, has become an indispensable step in the language tools' production and in the process of language computerization.

In the aim to promote the Amazigh language and to encourage its computerization, we have lead to build a morphosyntactic corpus; which has elicited the development of a tool, providing support and linguists' assistance.

Keywords – Mots Clés

Etiquetage morphosyntaxique, Outil d'assistance, Corpus, Langue Amazighe.

Part-of-Speech tagging, Assistance tools, Corpora, Amazigh language.

1 Introduction

Les corpus sont des outils indispensables et précieux en traitement automatique du langage naturel. Dans la littérature, plusieurs catégories de corpus sont distinguées à savoir : les *corpus bruts*, constitués de textes écrits ou de transcriptions écrites de productions orales ; les *corpus annotés* ou *enrichis* avec des indications relatives à la structure du texte, aux

catégories morphosyntaxiques ou sémantiques; les *corpus alignés* avec un contenu traduit en plusieurs langues. L'utilisation d'une catégorie ou d'une autre dépend principalement de l'axe de recherche et de l'application visés ; néanmoins, la disposition de toutes ces catégories est une étape préliminaire pour l'informatisation de la langue Amazighe.

Certes, les corpus annotés sont plus intéressants que les données brutes, cependant le coût du processus d'annotation en termes de ressources humaines en limite la quantité et la disponibilité. Ce coût est particulièrement important lorsque les corpus annotés satisfassent un certain nombre de critères de qualité, qui assurent que l'information ajoutée ait réellement une valeur. Comme dans tous les domaines où les masses de données à manipuler sont importantes, il est naturel que les spécialistes soient à la recherche des outils d'assistance, qui permettent de minimiser les efforts nécessaires pour produire des corpus et améliorer leurs qualités en permettant des vérifications et en simplifiant les modifications et les mises à jour.

C'est sous ces perspectives que l'Institut Royale de la Culture Amazighe a encouragé le développement de tels outils et la construction des ressources linguistiques, en particulier ceux liées à l'annotation morphosyntaxique.

Dans ce papier, nous présentons quelques caractéristiques de la langue Amazighe. Nous discutons de la construction des jeux d'étiquettes morphosyntaxiques, en considérant les spécificités linguistiques de cette langue. Et finalement, avant de conclure, nous décrivons la structure de l'outil d'assistance que nous avons développé.

2 Caractéristiques de la langue Amazighe

La langue Amazighe, connue aussi par le berbère ou tamazight, est considérée comme la langue "autochtone" de l'Afrique du Nord. Elle couvre toute l'Afrique du Nord, le Sahara et une partie du Sahel ouest africain. Mais les pays principalement concernés sont, par ordre d'importance démographique : le Maroc avec 50% de la population globale (Boukous, 1995), l'Algérie avec 25% de la population, le Niger et le Mali (Chaker, 2003).

Au Maroc, l'amazighe se répartit en trois grandes régions dialectales qui couvrent l'ensemble des régions montagneuses : au nord-est, le Rif avec le dialecte Tarifite ; au centre, le Moyen-Atlas et une partie du Haut-Atlas avec le dialecte Tamazighte ; au sud et sud-ouest, le Haut-Atlas, l'Anti-Atlas et Sous, le domaine chleuh avec le dialecte Tachelhite.

2.1 Alphabet Amazighe

Depuis l'Antiquité la langue Amazighe possède sa propre écriture, Tifinaghe, qui est toujours utilisée chez les Amazighes des zones sahariennes (Touaregs). Néanmoins, le long de l'histoire, particulièrement depuis la fin des années 1960, cette écriture a assimilé des changements et des développements dont l'objectif est de fournir à la langue Amazighe un système alphabétique standard plus adéquat et utilisable pour tous les parlers amazighes actuels. Ainsi en 2003, partant d'un héritage aussi bien ancien que moderne et contemporain, l'IRCAM a développé un système d'alphabet sous le nom de Tifinaghe-Ircam.

L'alphabet standardisé par l'IRCAM est basé sur un système graphique à tendance phonologique. Ce système ne retient pas toutes les réalisations phonétiques produites, mais uniquement celles qui sont fonctionnelles (cf. (Ameur et al., 2004)).

Cet alphabet comporte :

- 27 consonnes dont : les labiales (ⵃ, ⵈ, ⴱ), les dentales (ⵜ, ⴷ, ⵟ, ⴹ, ⵍ, ⵏ, ⵇ, ⵔ), les alvéolaires (ⵙ, ⵣ, ⵚ, ⵥ), les palatales (ⵛ, ⵊ), les vélaires (ⴽ, ⵅ), les labiovélaires (ⴽⷯ, ⵅⷯ), les uvulaires (ⵇ, ⵅ, ⵖ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⴰ) ;
- 2 semi-consonnes : ⵢ et ⵡ ;
- 4 voyelles : trois voyelles pleines ⴰ, ⵉ, ⵓ et la voyelle neutre (ou schwa) ⴻ qui a un statut assez particulier en phonologie amazighe.

2.2 Mot graphique

Parmi les diverses définitions du mot, les linguistes de l'IRCAM ont considéré un mot graphique comme une séquence de lettre, éventuellement une seule lettre, délimitée par deux blancs. Les groupements constituant un mot graphique en amazighe standard sont (Ameur et al., 2004) (Boukhris et al., 2008) :

- Le substantif avec ses marques de genre, de nombre et d'état ;
- L'adjectif avec ses marques de genre, de nombre et d'état ;
- Le verbe avec ses morphèmes dérivationnels (causatif, réciproque et passif), ses marques d'aspect (préfixées et infixées), et ses marques d'accord (genre, nombre et personne) ;
- Le participe avec ses morphèmes dérivationnels (causatif, réciproque et passif), ses marques d'aspect (préfixées et infixées), et ses marques d'accord (genre, nombre) ;
- Les pronoms objets direct et indirect ;
- Les démonstratifs de proximité, d'éloignement et d'absence ;
- Tout syntagme prépositionnel où le régime de la préposition est pronominal ;
- Tout blanc lexicalisé sur la base de l'adjectiveur ⵈⴻ, ⴱⴱⴻ;
- Tout blanc qui consiste en un quantificateur et son complément ;
- Les préverbes de négation et d'aspect ;
- Les prépositions ;
- Les pronoms autonomes ;
- Les adverbes ;
- Les conjonctions ;
- Les particules d'orientation ;
- La particule prédicative ;
- Les vocatifs ;
- Les présentatifs ;
- Les interrogatifs.

3 Annotation morphosyntaxique

L'étiquetage morphosyntaxique (Part-of-Speech tagging ou POS tagging en anglais) consiste à identifier pour chaque mot, à partir du contexte et des connaissances lexicales, une classe grammaticale, généralement définie selon un niveau de granularité.

L'annotation peut être vue comme la composition de trois fonctions, à savoir :

1. la segmentation du flux de caractères en mots ;
2. l'étiquetage a priori (hors-contexte) des mots au moyen des informations lexicales, qui associe toutes les étiquettes possibles pour un mot donné ;
3. la sélection, en fonction du contexte du mot, de l'étiquette la plus appropriée parmi celles identifiées par l'étiquetage a priori.

Ce qui rend cette tache très complexe, particulièrement lorsque l'étiquetage est manuel et la taille du corpus est représentative.

Afin de minimiser cette complexité et de réduire le coût du processus d'annotation en termes de ressources humaines, nous avons réalisé un outil d'assistance à l'étiquetage morphosyntaxique, spécifiquement pour la langue Amazighe. Cet outil permet d'automatiser l'ajout de l'information linguistique au corpus brut, la vérification et la correction des annotations, et la représentation du corpus annoté sous un format standard.

3.1 Jeu d'étiquettes

Les étiquettes généralement contiennent la partie du discours du mot concerné, accompagnée d'un certain nombre d'informations morphosyntaxiques (genre, nombre, temps, personne, etc.). Idéalement, un jeu d'étiquette doit permettre de :

1. représenter la richesse des informations lexicales,
2. représenter l'information nécessaire à la désambiguïsation en contexte des étiquettes morphosyntaxiques,
3. encoder les informations utiles au traitement linguistique pour lequel l'étiquetage morphosyntaxique a été déployé.

Des recommandations ont été formulées par EAGLES[1] pour la tache d'annotation morphosyntaxique. Elles portent principalement sur le choix des étiquettes et reposent sur une distinction entre étiquettes obligatoires, étiquettes recommandées et extension particulière.

Dans le cas de la langue Amazighe, la question de la classification des catégories grammaticales est une tâche difficile et toujours en débat au sein du centre de l'aménagement linguistique de l'IRCAM ; essentiellement du fait que cette langue est en cours de standardisation, et que ses règles sont en phase de construction. A ce jour, il n'existe aucun standard reconnu pour les catégories des mots amazighes. Pour cette raison, nous allons essayer de définir des jeux d'étiquettes qui seront utilisés dans les différentes applications du traitement automatique de l'Amazighe, et en particulier dans l'outil d'assistance à l'étiquetage, en attendant la version standard définitive qui sera arrêtée par les linguistes.

[1] http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html. Consulté en septembre 2009

1.	N [nom]	2.	V [verbe]	3.	AJ [adjectif]
4.	P [pronom]	5.	D [déterminant]	6.	AV [adverbe]
7.	PP [préposition]	8.	C [conjonction]	9.	PR [particule]
10.	I [interjection]	11.	F [focalisateur]	12.	NR [nombre]
13.	R [résiduel]	14.	SYM [symbole]	15.	PU [ponctuation]

Où, l'étiquette « résiduel » est assignée aux termes qui ne rentrent pas dans les catégories usuelles tels que les mots étrangers.

Figure 1 : Liste des étiquettes obligatoires

Ainsi, nous avons proposé deux listes d'étiquettes spécifiques aux caractéristiques de la langue Amazighe. Une contient les étiquettes obligatoires assignées lors de l'annotation morphosyntaxique (cf. le tableau de la Figure 1). L'autre contenant des étiquettes recommandées, qui fournissent des indications plus précises tels que le nombre et le genre pour les noms et les adjectifs, le temps, le mode et la personne pour les verbes.

Ci-dessous, nous présentons ce jeu des étiquettes en détaillant les subdivisions de chacune de ses catégories :

- Nom : En plus des attributs de genre (masculin, féminin) et de nombre (singulier, pluriel, collectif), d'autres attributs ont été définis : commun, propre, de parenté, numéral, non composé ; ainsi que l'attribut de l'état (libre ou construit).
- Verbe : Les catégories du verbe sont répertoriées sous forme de cinq sous classes :
  - Genre : masculin, féminin ;
  - Nombre : singulier, pluriel ;
  - Personne : 1ère, 2ème, 3ème personne ;
  - Mode : impératif, non impératif, participe ;
  - Aspect : aoriste, accompli, accompli négatif, inaccompli.
- Adjectif : Les attributs des adjectifs pour l'Amazighe sont :
  - Genre : masculin, féminin ;
  - Nombre : singulier ou pluriel ;
  - Etat : libre ou d'annexion.
- Pronom : L'attribut principal de cette catégorie est le type, et non pas le genre ou le nombre. L'ensemble des valeurs appropriées à cet attribut est : le démonstratif, le possessif, l'interrogative, le relatif, l'autonome, le régime direct, le régime indirect et l'indéfini.
- Déterminant : Les valeurs de l'attribut de cette classe sont l'indéfini, le démonstratif, le quantitatif, l'altérité, le numéral, le présentatif, le possessif, l'interrogatif et le prédicateur.
- Adverbe : Les adverbes de l'Amazighe sont utilisés essentiellement pour exprimer le temps, le lieu, la quantité, la cause et la manière. Ils peuvent être soit sous forme affirmative ou interrogative.

- Préposition : Les attributs de cette catégorie sont : le lieu, la direction, le temps, l'instrument, la possession et l'appartenance, l'accompagnement, et l'attribution. Ils peuvent être soit sous une forme libre ou affixée d'un pronom.
- Conjonction : Cette catégorie emploie l'attribut type avec deux valeurs :
  - Coordination qui peut être soit d'accompagnement, d'alternance, de disjonction ou cumulatif ;
  - Subordination qui peut accepter les attributs de temps, de but, de cause, de concession/opposition, de condition, de conséquence, de comparaison, de manière, de relativité ou de complétion.
- Particule : Sept attributs sont associés à cette classe : l'aspectuelle, l'orientation, la prédication, la négation, l'interrogation, la modale et la vocative.
- Interjection; Focalisateur; Nombre; Résiduel; Symbole; Ponctuation : Aucun attribut n'est associé à ces six catégories obligatoires.

Ce jeu d'attribut suivra quelques modifications une fois la liste des catégories grammaticales sera arrêtée par les linguistes.

3.2 Outil d'assistance

Dans l'objectif de simplifier l'annotation morphosyntaxique aux linguistes, nous avons développé un outil d'assistance, après une étude préalable des caractéristiques de la langue Amazighe. Ce dernier permet d'annoter les textes avec le jeu d'étiquettes recommandées, de saisir le lemme adéquat, de modifier les informations fournies, de vérifier la cohérence des informations annotées par l'ensemble des linguistes chargé de l'étiquetage et les mettre à jour, et de fusionner à la fin ces informations dans le corpus selon un format standard.

La structuration de cet outil est conçue en trois parties :

La première partie consiste en une segmentation des textes en phrases puis en mots isolés. Elle donne la possibilité de sélectionner pour chaque mot son étiquette grammaticale adéquate et de saisir le lemme correspondant. Comme elle permet de stocker les informations annotées dans une base de données.

La deuxième partie consiste à comparer les informations annotées par l'ensemble des linguistes travaillant sur le même document, à extraire les discordances existantes entre les annotations des différents linguistes, et à permettre les corrections et les mises à jour de ces annotations.

Tandis que la troisième partie, elle, concerne la mise en forme des annotations ainsi que le contenu textuelle de notre corpus sous un format standardisé.

4 Conclusion

Nous avons proposé deux jeux d'étiquettes grammaticales de la langue Amazighe, dont le premier correspond à des étiquettes obligatoires et le deuxième à des étiquettes recommandées. Puis, nous avons présenté la structure d'un outil d'assistance à l'annotation morphosyntaxique que nous avons proposé pour assurer de l'aide aux linguistes.

Le développement de cet outil ne présente qu'une étape préliminaire, dans un premier temps, dans le processus de la constitution d'un corpus annoté morphosyntaxiquement ; puis dans l'élaboration, dans un deuxième temps, d'un étiqueteur automatique, d'un lemmatiseur, d'un analyseur syntaxique et d'un extracteur terminologique.

Références

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. M., Souifi H. (2004), *Initiation à la langue amazighe*, Rabat, IRCAM.
Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008), *La nouvelle grammaire de l'amazighe*, Rabat, IRCAM.
Boukous A. (1995), *Société, langues et cultures au Maroc : Enjeux symboliques*, Casablanca, Najah El Jadida.
Chaker S. (2003), Le berbère, Actes de *Les langues de France*, 215-227.