Étiquetage morpho-syntaxique de textes kabyles

Sinikka LOIKKANEN
Université d'Helsinki
sinikka.loikkanen@helsinki.fi

Résumé. Cet article présente la construction d'un étiqueteur morpho-syntaxique développé pour annoter un corpus de textes kabyles (1 million de mots). Au sein de notre projet, un étiqueteur morpho-syntaxique a été développé et implémenté. Ceci inclut un analyseur morphologique ainsi que l'ensemble de règles de désambiguïsation qui se basent sur l'approche supervisée à base de règles. Pour effectuer le marquage, un jeu d'étiquettes morpho-syntaxiques pour le kabyle est proposé. Les résultats préliminaires sont très encourageants. Nous obtenons un taux d'étiquetage réussi autour de 97 % des textes en prose.

Abstract. This paper describes the construction of a morpho-syntactic tagger developed to annotate our Kabyle text corpus (1 million words). Within our project, a part-of-speech tagger has been developed and implemented. That includes a morphological analyser and a set of disambiguation rules based on supervised rule-based tagging. To realise the annotation, a POS tagset for Kabyle is proposed. The first results of tests are very encouraging. At present stage, our tagger reaches 97 % of success.

Mots-clés : Étiquetage morpho-syntaxique, corpus de textes, langue kabyle, berbère.

Keywords: Part of speech tagging, text corpus, kabyle language, berber.

1 Introduction

L'étiquetage morpho-syntaxique automatique est une technologie relativement bien maîtrisée. Au moins pour les langues européennes comme l'anglais qui déjà dispose de grands volumes de corpus étiquetés. L'étiquetage morpho-syntaxique ou grammatical consiste à affecter à chaque occurrence d'un corpus un symbole représentant sa catégorie grammaticale (nom, verbe, ...) et, éventuellement, les informations morphologiques associées (masculin, singulier, ...) (Paroubek & Rajman, 2000). Les méthodes et les applications conçues pour une langue ne sont pas telles quelles transférables aux autres langues. Mais au niveau de l'annotation grammaticale, il y a eu d'importants efforts d'harmonisation et de standardisation aux seins de projets internationaux tels que MULTEXT (MUL, 1996) et EAGLES (EAG, 1996). Les recommandations et les standards de ces projets rendent possible la comparabilité des différents corpus étiquetés en plusieurs langues.

Dans le cadre de notre projet, nous avons constitué un corpus de textes kabyles, CKL (corpus kabyle littéraire). Le fond du corpus est constitué des six premiers romans kabyles publiés entre 1981–1995 dont le vocabulaire est présenté dans notre mémoire de DEA (Loikkanen, 1998). Par la suite, le corpus a été progressivement complété au fur et à mesure de la disponibilité

de nouveaux textes. Le CKL comprend actuellement soixante textes intégraux qui représentent un million d'occurrences. Les textes représentent différents genres littéraires (romans, contes, poèmes, chansons, récits) et couvrent différents thèmes. Les romans représentent la moitié des occurrences bien qu'ils ne représentent que 20 % des textes. Les textes sont numérisés, la structure balisée en XML[1] selon les principes de TEI[2] et la transcription est marquée par Unicode[3].

Le kabyle est une langue pour laquelle le travail d'étiquetage est une tâche nouvelle. Le kabyle est une variante de la langue berbère parlé en Kabylie, en Algérie du Nord. Le berbère ou *tamazight* (nom berbère de langue) est une langue afro-asiatique (ou chamito-sémitique) qui est considérée comme la langue autochtone de l'Afrique du Nord (Camps, 1987).

Pour annoter notre corpus, nous avons conçu et développé un étiqueteur ainsi qu'un jeu d'étiquettes morpho-syntaxiques. Dans ce qui suit, nous présentons d'abord notre jeu d'étiquettes morpho-syntaxiques pour le kabyle. Puis, nous présentons les principes de l'analyse. Pour finir, les premiers résultats de cette annotation seront rapportés.

2 Jeu d'étiquettes morpho-syntaxiques pour le kabyle

Pour annoter notre corpus, nous avons élaboré un jeu d'étiquettes morpho-syntaxiques pour le kabyle. Actuellement, il n'existe pas, à notre connaissance, d'études disponibles ni de catalogues avec un classement morpho-syntaxique fin pour le kabyle ou pour les autres dialectes du berbère. Dans les grammaires du kabyle (Mammeri, 1986; Mammeri, 1988; Naït-Zerrad, 2001), on trouve les descriptions pour les parties du discours comme nom, verbe, adjectif, adverbe, pronom, préposition, conjonction, etc. ; on a appliqué tel quel au berbère les catégories grammaticales que l'on retrouvait, par exemple, dans les grammaires traditionnelles du français. En fait, l'organisation des classes en berbère n'est pas radicalement différente de celle que l'ont peut rencontrer dans les langues indo-européennes (Chaker, 1984). Si l'on néglige un certain nombre de phénomènes secondaires et de formes isolées, les unités syntaxiques du berbère s'organisent en quatre grands ensembles qui sont 1) les verbes, 2) les noms, 3) les connecteurs ou relationnels et 4) les déterminants divers. Les deux premiers sont des catégories lexicales, les deux derniers des catégories grammaticales. Sous la classe de connecteurs, on regroupe tous les indicateurs de relation, prépositions, subordonnants, conjonctions et connecteurs divers. La classe des déterminants forme un ensemble hétérogène où on peut isoler, entre autres, une sous-catégorie comme les adverbes (Chaker, 1984).

Pour construire notre jeu d'étiquettes, nous avons défini un classement aussi fonctionnel que possible. Le jeu d'étiquettes s'inspire du projet EAGLES (EAG, 1996). Bien que ces recommandations ne soient pas applicables telles quelles au kabyle, elles servent de point de départ pour s'orienter vers des applications multilingues ainsi que la réutilisabilité et la comparabilité d'étiquettes. Pour l'instant, notre jeu d'étiquettes contient 12 catégories principales : nom (N), verbe (V), adjectif (A), pronom (PR), adverbe (ADV), préposition (PREP), conjonction (C), numéral (NU), interjection (I), particule (P), résidu (R) et ponctuation (PU). Ces catégories se basent sur les parties du discours et sur les traits morphologiques décrits dans les grammaires kabyles. Le jeu d'étiquetage est présenté par les paires attribut–valeur, par exemple un nom commun féminin singulier à l'état libre, comme *taqcict* 'une fille', est codé de la façon suivante :

[1] Extensible Markup Language, http ://www.w3.org/XML/
[2] Text Encoding Initiative, http ://www.tei-c.org/
[3] http ://unicode.org/

`N[type=commun genre=féminin nombre=singulier état=libre]` ou `NCFSL`.

– **Nom** : Le nom kabyle varie en genre, en nombre et en état[4]. Pour le nombre, outre les formes du singulier et du pluriel, il existe une forme duelle empruntée à l'arabe, mais très peu utilisée et on ne l'emploie que lorsqu'il s'agit de temps (*cehṛayen* 'deux mois'). L'état se manifeste par un changement d'une voyelle initiale et, éventuellement, par un ajout d'une semi-consonne (*w*, *y*) au début d'un mot pour les masculins commençant par une voyelle, par exemple *axxam* (état libre) 'maison' devient une forme *wexxam* (état d'annexion). Mais, il existe une partie importante de mots invariables en état (les noms de parenté, les emprunts au français et à l'arabe), i.e. les mots n'ayant pas de modifications dans la forme et dont l'état n'est pas ainsi identifiable automatiquement hors contexte. Pour le nom kabyle, nous avons ainsi défini les attibuts suivants dont les valeurs sont présentées entre crochets :

  type [commun, propre], genre [masculin, féminin], nombre [singulier, pluriel, duel],
  état [libre, annexion, non-marqué].

– **Verbe** : Les verbes kabyles sont flexionnels et varient en genre, en nombre et en personne. Il existe trois aspects : l'aoriste, l'aoriste intensif (inaccompli) et le prétérit (accompli), ainsi que trois modes : l'indicatif, l'impératif et le participe. L'infinitif n'existe pas au sens où on l'entend par exemple en français ; le lemme ou la forme lexicale d'un verbe donné dans les dictionnaires est la forme à l'impératif de la 2^{e} personne du singulier. Notre proposition pour les paires attribut–valeur pour les verbes kabyles est la suivante :

  mode [indicatif, impératif, participe], thème [aoriste, aoriste intensif, prétérit],
  degré [positif, négatif], nombre [singulier, pluriel], personne [1re, 2^{e}, 3^{e}],
  genre [masculin, féminin, commun].

– **Adjectif** : L'adjectif kabyle se forme principalement sur les verbes d'état (*aberkan* 'noir' du verbe *ibrik* 'être noir') ; par son type, il est qualificatif. Les formes secondaires sont les formes empruntées à l'arabe, les formes invariables et les formes complexes (Chaker, 1995). L'adjectif partage toutes les caractéristiques morphologiques du nom, il varie en genre, en nombre et en état, sauf les invariables. Les degrés de comparaison sont indiqués avec les verbes ou avec les prépositions, il n'y a pas de formes graphiques pour le comparatif ou pour le superlatif de type *bon – meilleur*. Les paires attribut–valeur proposées sont les suivantes :

  genre [masculin, féminin, commun], nombre [singulier, pluriel],
  état [libre, annexion, non-marqué].

– **Pronom** : Dans cette catégorie, nous avons regroupé différents types de pronoms et de déterminants, comme les pronoms personnels et les pronoms démonstratifs. Les pronoms personnels varient en genre, en nombre et en personne. Ils sont autonomes (*netta* 'il, lui') ou des affixes (*-k* 'à toi') liés à un élément comme nom, verbe, préposition, adverbe ou présentatif (l'affixe du nom exprime la possession, l'affixe du verbe marque le complément direct ou indirect, l'affixe des prépositions et l'affixe des adverbes joignent l'état ou l'action à une personne). Les démonstratifs forment un groupe déterminatif spécial. Il s'agit du groupe des unités fréquentes et largement cultivées dans le parlé indiquant la situation d'un objet par rapport au locuteur. Ils sont autonomes ou suffixés au noms ; les premiers varient en genre et en nombre, les derniers sont invariables. Les paires proposées sont les suivantes :

  type [personnel, démonstratif, indéfini, interrogatif, relatif],
  position [autonome, affixe], pers-type [nom, préposition, verbe],

[4]État : Il s'agit d'un concept grammatical représentant un couple oppositif état libre / état d'annexion dans lequel la forme du nom change. L'état d'annexion marque la dépendance du nom par rapport aux autres éléments de la phrase (Chaker, 1995).

nombre [singulier, pluriel], personne [1re, 2e, 3e], genre [masculin, féminin, commun], location [proximité, éloignement, absence].

- **Adverbe** : Mot invariable, sauf quelques exceptions qui portent la forme nominale et qui varient en état. Par son type, l'adverbe décrit le temps (*azekka* 'demain'), la manière (*baṭel* 'gratuitement'), la quantité (*mliḥ* 'beaucoup') ou le lieu (*beṛṛa* 'dehors').
- **Préposition** : Les prépositions sont invariables, utilisées isolément devant un nom (*zdat wexxam* 'devant la maison') ou avec les affixes personnels (*zdat-i* 'devant moi').
- **Conjonction** : Elles sont de types de coordination (*ihi* 'donc', *walakin* 'mais') et de subordination (*mi* 'quand', *qbel* 'avant que'). Certaines d'elles sont utilisées dans les deux cas.
- **Numéral** : Ils sont de types cardinal (*tlata* 'trois') ou ordinal (*wis tlata* 'troisième'). Ils varient en genre sauf les numéraux empruntés à l'arabe qui ne s'accordent pas en kabyle.
- **Particule** : Elles sont de petits mots invariables servant à préciser le sens d'autres mots. Dans les recommandations EAGLES, certains de ces éléments sont regroupés sous la catégorie 'unique/unassigned'. Les particules sont : particule de l'aoriste (*ad*), particule prédicative (*d*), particule de négation (*ur*), particule complétive de négation (*ara*), particule vocative (*a*), particule exclamative (*ack-*), particule présentative (*aql-*) et particule d'orientation (*d*, *n*).

3 Le problème de la variabilité des transcriptions

Les textes du CKL datent de différentes décennies, les premiers viennent de la fin du 19e siècle, les plus récents sont de l'année 2006. Bien que les principes de la notation (PNB, 1996; ALB, 1998) soient bien établis, on trouve dans la pratique des transcriptions différentes.

- Au niveau des lettres les variations se manifestent en différents choix graphiques :

  (1) *eçç* = *ečč* 'manger', *ḥemmel* = *ẖemmel* 'aimer', *ḍleb* = *ḏleb* 'demander', *imjuhad* = *imǧuhad* 'combattant', *taabbuṭ* = *taâbbuṭ* = *taɛbbuṭ* 'ventre'.

- Au niveau des mots, il y a des hésitations concernant le nombre des consonnes à écrire :

  (2) *claɣem* = *cclaɣem* 'moustache', *tagara* = *taggara* 'moment', *tidet* = *tidett* 'vérité',

  le placement d'un point emphatique :

  (3) *aḍaṛ* = *aḍar* = *adaṛ* 'pied', *aṣaḍuf* = *asaḍuf* 'loi', *lbaruḍ* = *lbaṛud* 'poudre à canon',

  ou la notation de la labio-vélarisation des consonnes vélaires et labiales :

  (4) *ameqran* = *ameᵠran* = *ameq°ran* = *ameqʷran* = *ameqwran* 'grand'.

- Au niveau des phrases, les variations se manifestent par l'assimilation phonétiques aux frontières des mots :

  (5) *ţ-ţideţ* ← *d tideţ*, *a d awi* ← *a d-tawi*.

  ou par la présence ou l'absence d'un tiret entre le nom ou le verbe et leurs affixes :

  (6) *yellis* = *yelli-s* = *yell-is* = *yelli s* 'sa fille'.

Finalement, la position du ə-muet pose certains problèmes. Il s'agit d'une voyelle neutre qui n'a pas de statut phonologique et dont la situation dépend de ce qui l'entoure. Par convention il est noté par une lettre 'e' dans tous les textes kabyles à quelques exceptions près. On ne peut pas nier son existence, au niveau graphique sa présence est importante : 12 % des lettres parmi toutes les lettres dans les six premiers romans kabyles (le deuxième rang après la lettre 'a'). Le placement de cet 'e' varie dû à l'enchaînement des mots :

(7) *iger* 'il a mis', mais *yegr-as* 'il lui a mis' au lieu d'écrire *iger-as*.

Pour gérer toutes ces variations et pour annoter le corpus, nous avons développé un analyseur qui se base sur les expressions régulières dans la reconnaissance des mots. Ainsi, nous trouvons des occurrences qu'elles soient écrites avec une ou plusieurs consonnes similaires au sein d'un mot (c{1,}laɣem), avec ou sans point sous la lettre (a(d|ḍ)a(r|ṛ)) ou qu'elles soient écrites différemment à cause de l'assimilation, comme pour le mot *abrid* 'chemin' à l'état d'annexion :

(8) [g] *brid* — [deǧǧ]-*ebrid* — [deg]-*gwebrid* — [deg]-*webrid* — [f]-*febrid* — [anebdu] *bwebrid* — [tikli] *bbwebrid* — *wubrid* — *ubrid.*

Le problème de la variabilité des transcriptions a ces conséquences aussi dans la segmentation des mots. La segmentation par des blancs n'est pas suffisante. Pour avoir des occurrences analogues, les unités textuelles sont ou découpées ou regroupées ou bien les deux à la fois :

– Les chaînes sont découpées lorsqu'il s'agit des affixes reliés par un trait d'union au mot auquel ils se rapportent. Parfois les mêmes unités sont écrites ensemble sans tiret entre les composants. Par exemple, les compositions *yemma s* | *yemma-s* | *yemmas* 'sa mère'.
– Les unités sont regroupées lorsqu'il s'agit des noms composées (*adrar ufud* 'tibia (montagne de jambe)', *Ait Ahmed*) ou les numéraux composés (*xems meyya* 'cinq cents'). Un grand groupe causant des hésitations dans l'écriture sont les prépositions complexes ; nous avons réuni les composants séparés par des blancs, par exemple, pour *seddaw* 'en-dessous de', nous avons aussi les formes *s ddaw* | *s eddaw* | *s-eddaw* | *si ddaw*.
– Parfois, les chaînes sont reconstruites. Par exemple, la composition comme *aqli-y-i* 'me voici' est découpée aux frontières des mots, puis, le reste a été regroupé pour avoir des occurrences analogues (*aql*|| *i-y-i*) à une forme « canoniques » *aql-iyi* 'me voici'.

4 Analyse morphologique

À cause de la transcription très instable, l'énumération de toutes les formes dans un lexique est une tâche impossible. En fait, nous n'avons pu utiliser les listes préalablement établies que pour les mots de classes grammaticales. Pour les autres classes, notamment pour les verbes et les noms, nous avons adopté une autre approche.

Bien que le kabyle soit une langue où la construction des mots se base principalement sur l'usage des racines, l'analyse se basant sur le modèle « interdigitation » de style sémitique n'est pas applicable telle quelle. Premièrement, parce que l'alternance vocalique n'est pas schématique ni toujours possible à prévoir, et deuxièmement, les racines kabyles ne portent pas un sens unique par racine. En fait, il existe plusieurs racines avec plusieurs champs sémantiques distincts, par exemple, la racine $\sqrt{br}$ a une vingtaine de sens différents.

Pour notre analyseur, nous avons fixé les exigences suivantes :

1. La langue que nous allons analyser se base sur l'utilisation des racines ; l'analyse doit produire de chaque forme au minimum un lemme et les consonnes de la racine.
2. L'analyseur doit identifier toutes les formes avec toutes les transcriptions possibles et rendre les descriptions morphologiques.
3. Pour identifier les formes, l'analyse se base sur l'extraction des stemmes, à partir desquels on dérive les lemmes et les racines.
4. Les dépendances de longues distances doivent être respectées, seulement les couples d'affixe « légaux » seront acceptés.

Par les couples d'affixe « légaux » on comprend les paires préfixes–suffixes correctes qui forment les mots. Par exemple, pour les verbes ordinaires nous avons 9 formes conjuguées en personne, 3 formes pour l'impératif et 5 formes de participe. Parmi ces formes, trois préfixes (∅-, `t-` et `i-`, dont le dernier peut se manifester aussi comme `y-`) et 3 suffixes (`-mt`, `-n` et -∅) sont ambigus. Les couples légaux sont les paires qui se réalisent correctement.

Pour construire un analyseur, on pourrait bien former trois listes dont la première pour les préfixes, la deuxième pour les stemmes et la troisième pour les suffixes, et construire un automate sur ces trois listes. Le problème dans cette approche est ce que le nombre de stemmes différents peut devenir relativement grand à cause des différentes notations utilisées dont nous ignorons a priori la nature des variations. Notre solution est l'identification des composants en expressions régulières. Initialement, nous n'avons les listes numériques que pour les affixes, pas celles de stemmes ou de racines. Ces dernières sont ici dérivées à partir des données du corpus. Notre analyseur identifie d'abord les composantes morphologiques des formes données, puis, à partir des stemmes obtenus, l'identification des racines et des lemmes est effectuée. Par exemple, pour la forme prétérit d'un verbe comme *turgamt* 'vous (*f.*) avez rêvé' la base de l'analyse est le stemme *urga* (racine : *rg* ; vocalisation : *u-a* ; stemme : *urga* 'prétérit' ; affixes : *t-, -mt* ; lemme : *argu* 'rêver') auquel on ajoute les affixes par la concaténation. Pour extraire ces composants, une règle est définie :

```
if ($w =~ /(^t)(.*?)mt$/) { $stemme = $2; }
```

Cette règle veut dire que si la forme commence par `t-` et finisse par `-mt`, la forme est vraisemblablement une forme verbale de la 2e personne féminine du pluriel. Ce qui reste, dans le milieu dans la variable $2, est le stemme avec toutes les transcriptions possibles. À partir de ce stemme nous obtenons les consonnes de la racine, le lemme et l'information concernant l'aspect verbal. Ce n'est que dans cette phase que l'on recourt à l'aide des dictionnaires pour définir la forme du lemme. Par exemple, pour les motifs *tegremt* et *tersemt*, pour lesquels on obtient les stemmes *egre* et *erse* respectivement, on défini ainsi avec les dictionnaires les formes du lemme qui sont pour ces verbes *ger* 'mettre' et *ers* 'descendre ; se poser ; se calmer'.

5 Désambiguïsation

On s'appuie sur l'approche à base de règles. On tente de minimiser l'ambiguïté par l'analyse du contexte proche ; ce processus repose sur l'hypothèse que la catégorie d'un mot dépend d'un contexte local, i.e. de la catégorie d'un mot précédent ou suivant (fenêtre = 1). Les définitions se basent ou sur les catégories grammaticales ou sur les catégories grammaticales et les traits morphologiques. Les règles contextuelles à définir sont de types :

```
« si X est marqué NOM et VERBE, et que X est précédé d'un Y marqué
PREP, alors l'étiquette VERBE doit être supprimée de X »
```

```
« si X = 'd', et que X est suivi d'un Z marqué NOM+EA, alors
X = 'PREP' »
```

Ces règles reposent sur l'analyse des mots proches qui peuvent eux-mêmes être ambigus, dans ces cas, les règles fonctionnent en cascade. Prenons par exemple la forme *yesli* (figure 1) dans une chaîne comme *axxam n yesli* 'la maison du jeune marié' (exemple 9) dans laquelle nous avons deux mots avec deux interprétations :

(9)

axxam	*n*	*yesli*
NOM	X	X
maison	?	?

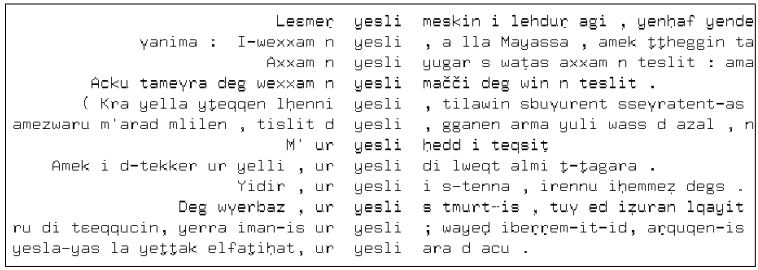

FIG. 1 – Début de la concordance de la forme *yesli* dans le CKL.

n : 1) préposition 'de, appartenant à' ; 2) particule d'orientation 'vers ici'.
yesli : 1) nom commun masculin singulier *isli* 'jeune marié', la forme à l'état d'annexion ; 2) prétérit négatif du verbe *sel* 'entendre', forme masculine 3e personne du singulier 'il n'a pas entendu'.

Les deux X sont résolus en deux phases : dans la première phase, on défini que *n* est une préposition s'il est précédé d'un nom ou d'un pronom et, dans la deuxième, on défini que *yesli* est un nom comme il a été précédé d'une préposition. Ces règles excluent les interprétations *n* = particule d'orientation et *yesli* = verbe qui se manifestent dans les phrases comme :

(10) *a n yuɣal* : particule de l'aoriste + particule d'orientation + verbe
ur yesli : particule de négation + verbe.

Dans certains cas, pour résoudre des ambiguïtés, la fenêtre a été étendue à couvrir les cas ± 2, et même plus, de manière à encadrer des classes grammaticales (affixes personnels, démonstratifs). Cela rend possible la résolution des cas comme :

(11) *axxam-nni n yesli* 'la maison (en question, dont on parle) du jeune marié'.

Les règles contextuelles se basant sur les catégories grammaticales ou sur les catégories grammaticales et les traits morphologiques ne peuvent pas toujours lever l'ambiguïté. Prenons comme exemple un mot kabyle très fréquent comme *d*, qui a, entre autres, les sens 1) particule prédicative 'c'est, ce sont' et 2) préposition 'et, avec'. Le sens de ce *d* peut être distingué, lorsqu'il précède un nom, par l'état de ce nom : si le nom qui suit est à l'état libre (EL), *d* a le sens 1 (exemple 12), si le nom qui suit est à l'état d'annexion (EA), *d* a le sens 2 (exemple 13).

(12) *argaz d aɣyul*
homme c'est âne+EL
l'homme est un âne

(13) *argaz d weɣyul*
homme et âne+EA
l'homme avec un âne

Mais, si le nom qui suit ne porte pas la marque de variation de l'état, comme par exemple dans l'énoncé *argaz d mmi-s* (exemple 14), nous avons deux sens selon le contexte : 1) l'homme, c'est son fils, 2) l'homme et son fils. Nous marquerons par EI (état invariable) les noms ne changeant pas de forme au début du mot.

(14) *argaz d mmi -s*
homme *d* fils+EI son
l'homme ? son fils

Dans ce dernier cas, pour enlever les ambiguïtés d'une façon automatique, on pensait à recourir aux méthodes statistiques de désambiguïsation, mais les premiers tests effectués ne confortent pas cette approche.

6 Tests et résultats

Nous avons testé notre système avec deux séries d'extraits de corpus. La première série a été utilisée comme corpus d'apprentissage pour la définitions des règles de désambiguïsation, la deuxième pour la validation. Les corpus de test ont été construits à partir des fragments de textes bruts de prose (contes, récits, romans). Pour la première série, on a pris de 6 textes différents les premiers 12000 bytes par texte. Cette quantité a puis été complétée pour terminer les dernières phrases obtenues jusqu'au point suivant terminant la phrase. Cela fait 1800–2500 mots par fragment (comptés par des blancs). Après la segmentation, le nombre se fixait entre 2000–2300.

Après l'analyse morphologique hors contexte, on constate qu'en moyenne un quart des mots était marqué comme ambigu. Sur la base de ces cas ambigus, on a écrit une centaine de règles de désambiguïsation. Les résultats sont vérifiés manuellement et, en cas d'erreurs, les règles sont corrigées, l'algorithme exécuté et le résultat vérifié. Finalement, on a obtenu 150 règles écrites. De cette manière on a construit itérativement un ensemble de règles de désambiguïsation. Après l'application de ces règles, il ne nous restait qu'en moyenne 1 % de cas ambigus. Le taux de réussite de la désambiguïsation est définit comme le pourcentage d'étiquettes assignées correctes par rapport au nombre total d'étiquettes assignées. Voir le tableau récapitulatif (tableau 1) où les résultats du premier tour avec les six fragments textuels.

1 N[a]	2 w[b]	3 <w>[c]	4 X S[d]	5 X S %[e]	6 X D[f]	7 %-final[g]
F1	1988	2314	623	26,9	31	98,6
F2	2514	2321	678	29,2	13	99,4
F3	2262	2189	488	22,3	23	98,9
F4	2064	2150	495	23,0	15	99,3
F5	2082	2052	382	18,6	8	99,6
F6	1803	2198	542	24,7	13	99,4

[a] Numéro de fragment.
[b] Nombre des mots w dans le texte brut, découpage par des blancs.
[c] Nombre des mots étiquetés en balise <w> ... </w> (hors ponctuation).
[d] Nombre des cas X, i.e. des mots ayant de multiples interprétations, dans les textes segmentés correctement après l'analyse morphologique, avant l'application des règles de désambiguïsation.
[e] Précédant en pourcentage.
[f] Nombre des cas X restant ambigus après l'application des règles de désambiguïsation.
[g] Pourcentage d'étiquettes réussites.

TAB. 1 – Résultats du premier tour.

Les cas restants encore ambigus, environ 1 %, sont les cas où $X = d$ et pour lesquels nous n'avons pas pu écrire des règles, par exemple, comme dans la phrase suivante (exemple 15) :

(15) *yekseb* *tameṭṭut* ***d*** *lall* *n* *nnif* ***d*** *lḥeṛma*
VPS3M NCFSL **X** NCFSI PREP NCMSI **X** NCFSI
il-a-possédé une-femme **?** propriétaire de honneur **?** respect
il avait une femme (**c'est** une propriétaire) d'honneur **et** de sacrée

Pour désambiguïser ces cas automatiquement, on a examiné les distributions des cas $X = d$ dans les fragments du corpus d'apprentissage en gardant à l'esprit les méthodes probabilistes de désambiguïsation. Les nombres obtenus étaient

$d = PO$ (particule d'orientation) : 50 %
$d = PD$ (particule prédicative) : 40 %
$d = PREP$ (préposition) : 10 %.

Parmi les cas qui restaient encore ambigus et que nous avons ensuite annotés manuellement, la distribution des cas $X = d$ est la suivante :

$d = PO$ (particule d'orientation) : 0 %
$d = PD$ (particule prédicative) : 60 %
$d = PREP$ (préposition) : 40 %.

Ces nombres indiquent que pour résoudre les derniers cas $X = d$ (+ N+EI, nom à l'état invariable) la résolution doit être cherchée ailleurs ; les méthodes probabilistes simples ne sont pas capables de les résoudre lorsqu'on obtient des répartitions presque égales (40 % / 60 %).

La deuxième série de test (tableau 2), validation des règles, a été effectuée avec deux fragments de texte plus larges. Pour cela, on a pris les premiers 50000 bytes de deux textes bruts en prose (un roman, un conte) qui ont ensuite été complétés pour terminer les dernières phrases obtenues jusqu'au point suivant terminant la phrase. Ces deux textes n'ont pas été utilisés dans la dérivation des règles de désambiguïsation (serie 1 ci-dessus). La validation est faite avec les règles définies dans le premier test. Cette fois, le taux de résolution du texte a atteint 97 %. En examinant les listes des cas restant non résolus, on constate qu'il y a 1) des cas où $X = d$ (+ N+EI), 2) des nouveaux cas X pour lesquels il n'y avait pas encore de règles ainsi que 3) des cas pour lesquels les règles déjà existaient mais qui restaient ambigus à cause de la transcription variable.

1 N[a]	2 w[b]	3 <w>[c]	4 X S[d]	5 X S %[e]	6 X D[f]	7 %-final[g]
T1	8289	9335	2169	23,2	162	98,2
T2	8074	9007	2055	22,8	239	97,3

[a−−g] Voir les explications ci-dessus, tableau 1.

TAB. 2 – Résultats du deuxième tour.

7 Conclusions et perspectives

Dans cet article, nous avons présenté notre travail sur l'étiquetage d'un corpus kabyle littéraire. L'étiquetage morpho-syntaxique est une tâche nouvelle dans le domaine du berbère. Il n'existe, à notre connaissance, aucun corpus kabyle, ni berbère, sous format numérique qui soit grammaticalement annoté, publié et disponible. Au sein de notre projet, nous avons construit

un étiqueteur, un jeu d'étiquettes morpho-syntaxiques ainsi qu'un formalisme pour résoudre les ambiguïtés dues aux homographes. Notre étiqueteur est constitué de trois modules : le segmenteur, l'analyseur et le désambiguïsateur. Dans toutes les étapes, les spécificités du corpus ont été prises en compte.

Les résultats préliminaires des tests effectués sont très encourageants. La désambiguïsation faite à base de règles a pu atteindre un taux d'étiquetage correct de 97 % des textes en prose.

La construction d'un étiqueteur morpho-syntaxique pour une langue pour laquelle les ressources numériques (lexiques, textes annotés) et les définitions formelles sont encore en cours de développement est un processus qui avance pas à pas. Avec notre méthode nous produisons à partir des données du corpus les ressources qui manquent : les premières listes numériques des stemmes, des lemmes et des racines qui peuvent être utilisées ultérieurement dans les futures versions de l'étiqueteur. Dans les perspectives d'évaluation du jeu d'étiquettes, une analyse plus profonde concernant les parties du discours et la catégorisation des mots grammaticaux est envisagée. La précision et l'extension des règles de désambiguïsation ainsi que de nouveaux tests se basant sur ces règles plus étendues avec des données sur une grande échelle sont aussi prévues.

Références

ALB (1998). *Aménagement linguistique de la langue berbère*. Inalco, Paris. `http://www.inalco.fr/crb/docs_pdf/amenage1998.pdf` [10.9.2006].

CAMPS G. (1987). *Les Berbères. Mémoire et identité*. Paris : Éditions Errance.

CHAKER S. (1984). *Textes en Linguistiques Berbère. Introduction au domaine berbère*. CNRS.

CHAKER S. (1995). *Linguistique berbère. Étude de syntaxe et de diachronie*. Paris : Peeters.

EAG (1996). EAGLES, Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R. `http://www.ilc.cnr.it/EAGLES96/home.html` [15.1.2007].

LOIKKANEN S. (1998). Vocabulaire du roman kabyle (1981–1995). Une étude quantitative. *Études et documents berbère*, **15–16**, 185–196.

MAMMERI M. (1986). *Précis de grammaire berbère (kabyle)*. Paris : AWAL.

MAMMERI M. (1988). *Tajeṛṛumt n tmaziɣt (tantala taqbaylit)*. Paris : AWAL/La Découverte.

MUL (1996). Multilingual Text Tools and Corpora. `http://www.lpl.univ-aix.fr/projects/multext` [15.1.2007].

NAÏT-ZERRAD K. (2001). *Grammaire moderne du kabyle. Tajerrumt tatrart n teqbaylit*. Paris : Karthala.

PAROUBEK P. & RAJMAN M. (2000). Étiquetage morpho-syntaxique. In J.-M. PIERREL, Ed., *Ingénierie des langues*, p. 131–150. Paris : HERMES Science Europe.

PNB (1996). *Propositions pour la notation usuelle à base latine du berbère*. Inalco, Paris. `http://www.inalco.fr/crb/docs_pdf/notation.pdf` [10.9.2006].