Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard

Delphine Bernhard[1] Amalia Todirascu[1] Fanny Martin[2] Pascale Erhart[1]
Lucie Steiblé[1] Dominique Huck[1] Christophe Rey[2]
(1) LiLPa - EA 1339, Université de Strasbourg
(2) Laboratoire Habiter le Monde - HM - EA 4287, Université de Picardie Jules Verne, Amiens

{dbernhard,todiras,pascale.erhart,lucie.steible,dominique.huck}@unistra.fr
{fanny.martin,christophe.rey}@u-picardie.fr

RÉSUMÉ

La tokénisation est une étape essentielle dans tout système de traitement automatique des langues, d'autant plus que de nombreux outils dépendent du découpage obtenu. La tâche est particulièrement ardue pour les textes qui ne respectent pas les conventions orthotypographiques ou les langues pour lesquelles ces conventions ne sont pas stabilisées. Nous nous intéressons ici aux cas de deux langues régionales de France, l'alsacien et le picard. Nous présentons les défis posés par ces deux langues, et proposons des critères de découpage implémentés dans des tokéniseurs.

ABSTRACT

Tokenization Issues for Two Regional Languages of France, Alsatian and Picard

Tokenization is an essential step in any language processing system, especially as many tools rely on the obtained segmentation. The task is particularly difficult for texts that do not meet typographical syntax or languages for which the usage of typographic signs is not stabilized. Here we focus on the case of two regional languages of France, Alsatian and Picard. We present the challenges of these two languages, and propose segmentation criteria implemented in tokenizers.

MOTS-CLÉS : Tokénisation, Alsacien, Picard, conventions orthotypographiques.

KEYWORDS: Tokenization, Alsatian, Picard, typographical syntax.

1 Introduction

La tokénisation est une des premières étapes de tout système de traitement automatique des langues (Webster & Kit, 1992). Cette tâche de découpage d'un texte en mots et phrases est donc d'une importance primordiale. Si la tokénisation pose encore rarement des problèmes pour les langues bien dotées, du moins pour les documents respectant les conventions orthotypographiques, la situation est toute autre pour de nombreuses langues moins dotées. Dans de nombreux cas il n'existe pas d'acte officiel de standardisation de la langue, y compris pour ce qui est de l'utilisation de l'espacement et des signes de ponctuation. C'est le cas notamment des deux langues régionales de France que nous considérons dans cet article : l'alsacien et le picard. Ces deux langues appartiennent à des familles différentes, mais posent des problèmes similaires, notamment les marques d'oralité à l'écrit (épenthèses - insertions de sons pour faciliter l'articulation, contractions signalées ou non par une

apostrophe). La tokénisation se base généralement sur des délimiteurs, qui marquent les frontières de mots et de phrases. Certains de ces délimiteurs sont non ambigus (comme le point d'exclamation, les double-points), d'autres sont ambigus (comme l'apostrophe, l'espace, le tiret ou le point) et nécessitent donc des traitements plus fins. Par exemple, en picard, l'apostrophe ne doit pas être considérée comme une frontière de mot dans la forme *k'min* (chemin) ou *batt'meints* (battements). On trouve le même phénomène en alsacien pour le déterminant *d'r* (le) ou le participe passé *g'hàlte* (arrêté). Il s'agit dans notre cas de développer des tokéniseurs afin de (i) produire des corpus annotés manuellement avec des informations morphosyntaxiques et (ii) développer des outils d'étiquetage morphosyntaxique pour ces deux langues régionales.

Nous passons tout d'abord en revue les travaux existants sur le découpage de textes en mots. Nous présentons ensuite les problèmes spécifiques qui se posent pour l'alsacien et le picard et les solutions proposées. Nous faisons ensuite une évaluation des systèmes de tokénisation développés.

2 État de l'art

La mise en œuvre d'un système de tokénisation passe dans un premier temps par la définition de ce qui constitue un token dans la langue considérée. Webster & Kit (1992) considèrent le token comme un "nœud terminal" (*terminal node*), qui, du point de vue des traitements ultérieurs, ne sera pas découpé en unités plus petites. Ainsi, un token pourra être constitué de plusieurs mots graphiques, comme c'est le cas par exemple des expressions idiomatiques, des mots composés (*pomme de terre*) ou de certains nombres (10 000). A l'inverse, un mot graphique peut être décomposé en deux unités (par exemple *au* décomposé en *à le*). Webster & Kit (1992) soulignent également l'importance de la tâche pour laquelle la tokénisation est effectuée, qui contraint la décomposition ou non des unités linguistiques en unités plus petites.

Les outils de TAL (outils de traitement de corpus, étiqueteurs, analyseurs syntaxiques) mettent en œuvre des stratégies de tokénisation plus ou moins évoluées même pour la langue standardisée. On distingue les approches à base de règles et les approches par apprentissage. Les premières définissent des règles de découpage ou d'identification des tokens qui sont, au moins en partie, dépendantes de la langue cible et qui peuvent prendre la forme d'expressions régulières (Grefenstette & Tapanainen, 1994). Les approches par apprentissage peuvent procéder par apprentissage supervisé, à l'aide d'un corpus pré-tokénisé, ou de manière non-supervisée. Par exemple, Jurish & Würzner (2013) utilisent un modèle de Markov caché (*Hidden Markov Model*) pour classifier les frontières de segments. L'approche non supervisée de Wrenn *et al.* (2007) repose quant à elle sur des arbres préfixes (*trie*) afin d'analyser les propriétés statistiques des frontières de tokens.

Enfin, les stratégies développées pour un type de textes à l'écrit peuvent ne pas fonctionner aussi bien pour d'autres types de textes pourtant dans la même langue. La problématique de la tokénisation resurgit ainsi pour les textes de spécialité, par exemple dans le domaine médical (Rabary *et al.*, 2015). Dans ce cas, le nombre très important de sigles et abréviations (comme *chir.* pour *chirurgie*) pose problème à un tokéniseur développé pour des textes journalistiques, de même que les mesures (comme *3x/j*). D'autres cas problématiques apparaissent dans les textes parus sur les médias sociaux, liés aux phénomènes de faute de frappe (espaces manquants, superflus ou mal placés), aux marques d'oralité (contractions) et aux répétitions de signes de ponctuation ou graphèmes, ainsi que tous les tokens spécifiques à ce type de communication comme les émoticônes ou mots-dièse (Laarmann-Quante & Dipper, 2016).

3 Tokénisation pour l'alsacien

3.1 Recommandations orthotypographiques existantes

Même s'il existe des propositions récentes de conventions orthographiques (par exemple ORTHAL (Zeidler & Crévenat-Werner, 2008)), l'écriture des dialectes alsaciens n'est pas strictement normée et les usages sont donc très diversifiés. Il est d'ailleurs très difficile des trouver des recommandations explicites pour l'usage des apostrophes ou des graphies séparées ou synthétiques, à part quelques exemples :

Grammaire de l'alsacien, d'Edmond Jung La grammaire formule des préconisations pour les adverbes pronominaux composés (*wo/da* + préposition), les articles et les pronoms personnels. Il est notamment conseillé de détacher les pronoms personnels sujets et objets dans tous les cas : *gìm mer s* (donne-le moi) et non *gìmmers* (Jung, 1983, p. 106-108).

Orthographe alsacienne, de Edgar Zeidler et Danielle Crévenat-Werner La méthode ORTHAL donne des recommandations très souples pour "l'écriture agglomérée". Ainsi, les graphies suivantes sont toutes considérées comme acceptables (Zeidler & Crévenat-Werner, 2008, p. 62) : *ìnere Stund* / *ìn ere Stund* / *ìn're Stund* (dans une heure). Ces diverses possibilités ne sont pas sans poser des problèmes à la tokénisation, car si l'on souhaitait avoir un découpage cohérent, il faudrait alors procéder à un découpage au sein de groupes de caractères alphanumériques contigus (comme par exemple découper *ìnere* en *ìn ere*). Nous avons dans ce cas fait le choix de ne procéder au découpage que si la séparation est marquée par une espace ou un signe spécifique (tiret et apostrophe). Par ailleurs, les recommandations données ci-dessus ne reflètent pas nécessairement les usages qui se rencontrent dans les données, pour une période qui s'étend sur 200 ans (1816 étant l'année de parution de la première pièce de théâtre en alsacien, le *Pfingstmontag* de Georges-Daniel Arnold).

3.2 Critères de découpage tokénisation en alsacien

Nous proposons de différencier les deux types de tokens suivants :

1. les signes graphiques qui relèvent du champ lexical ou grammatical ;
2. les signes graphiques qui relèvent de l'épenthèse, essentiellement le <n> et le <w> euphoniques. L'alsacien tend à représenter graphiquement des phénomènes phonétiques, ce pourquoi il a été choisi de découper les épenthèses, comme c'est souvent le cas dans le cadre de l'étiquetage de corpus oraux (Benzitoun *et al.*, 2012).

La Table 1 donne quelques exemples des deux types de phénomènes, relevés dans des textes variés et d'auteurs différents.

3.3 Développement d'un tokéniseur pour l'alsacien

Le tokéniseur développé repose sur des expressions régulières inspirées de (Grefenstette, 1998; Pointal, 2004). Les expressions régulières permettent de distinguer différents types d'unités :

– Les caractères et chaînes de caractères qui doivent être séparés de la suite lorsqu'ils sont situés après une espace : `d', s', z', n', üf'` etc. Cette règle permet de considérer les articles, prépositions et conjonctions élidés avant un mot comme des unités indépendantes.

Graphie initiale	Tokénisation proposée	Phénomène	Genre et source
zitter'm Ààfàng (depuis le début)	zitter␣'m Ààfàng	préposition + article au datif	encyclopédie : (Wikipedia, 2015)
in'ra Volkssproch (dans un dialecte)	in␣'ra Volkssproch	préposition + article au datif	encyclopédie : (Wikipedia, 2017)
uf'e'me Schàrebbà (sur un char à bancs)	uf␣'e'me Schàrebbà	préposition + article au datif	récit : (Sonnendrücker & Kauss, 1998)
hàt fànga-n-à drucka (a commencé à imprimer)	hàt fànga␣-n-␣à drucka	épenthèse	encyclopédie : (Wikipedia, 2017)
mine-n-Anforderunge (mes exigences)	mine␣-n-␣Anforderunge	épenthèse	théâtre : (Stoskopf, 1906)
Heere-n-Er (entendez-vous)	Heere␣-n-␣Er	épenthèse	théâtre : (Redslob, 1907)
geh-w-i (je vais)	geh␣-w-␣i	épenthèse	guide : (Keck & Daul, 2010)

TABLE 1 – Exemples de découpages en mots proposés pour l'alsacien.

– Les caractères et chaînes de caractères qui doivent être séparés de ce qui précède lorsqu'ils sont situés avant une espace : `'m'r, 'ma, 'me` etc. Cette règle permet notamment de considérer les pronoms situés à droite d'un verbe comme des unités lexicales indépendantes.
– Les suites de caractères qui doivent être séparés de ce qui suit lorsqu'ils sont situés en milieu de mot : `-n-`.
– Les nombres, abréviations, URLs, adresses mail.

La tokénisation a notamment pour objectif d'obtenir des unités cohérentes pour l'annotation morpho-syntaxique. Il est donc nécessaire de séparer les mots grammaticaux des mots relevant des classes ouvertes du lexique.

3.4 Évaluation du tokéniseur pour l'alsacien

Nous avons constitué un corpus de test à partir d'extraits appartenant à des genres variés, qui correspondent aux types d'écrits qu'il est possible de trouver en alsacien : théâtre (426 tokens), poésie (368 tokens), récit en prose (732 tokens), Facebook (333 tokens) et Wikipédia alémanique (778 tokens). Le corpus de test comprend 2 633 tokens en tout (signes de ponctuation inclus) et la tokénisation a été effectuée manuellement à partir des critères de découpage détaillés dans la section précédente. Nous avons également comparé les résultats de notre tokéniseur spécifique aux dialectes alsaciens avec le tokéniseur fourni avec le TreeTagger (Schmid, 1994), en prenant en compte le lexique des abréviations fournies pour l'allemand.

Les résultats de la tokénisation ont été évalués à l'aide de la commande `diff` et figurent dans les tables 2 et 3. Nous avons mesuré le nombre de vrais positifs (VP), faux positifs (FP, insertions par rapport à la tokénisation de référence) et faux négatifs (FN, absence de découpage par rapport à la référence), nous permettant ensuite de calculer la précision, le rappel et la F-mesure. Notre tokéniseur adapté à l'alsacien obtient des niveaux de performance supérieurs au tokéniseur générique fourni par le TreeTagger, même si les performances de ce dernier restent relativement bonnes. De manière globale, 2 621 tokens ont été correctement reconnus par notre tokéniseur, et 2 528 par le tokéniseur

du TreeTagger.

Genre	VP	FP	FN	Précision	Rappel	F-mesure
Facebook	332	1	1	0,997	0,997	0,997
Poésie	366	0	2	1,000	0,995	0,997
Récit	731	0	1	1,000	0,999	0,999
Théâtre	419	1	7	0,998	0,984	0,991
Wikipédia	773	1	5	0,999	0,994	0,996

TABLE 2 – Résultats de l'évaluation du tokéniseur spécifique aux dialectes alsaciens.

Genre	VP	FP	FN	Précision	Rappel	F-mesure
Facebook	325	1	8	0,997	0,976	0,986
Poésie	345	7	23	0,980	0,938	0,958
Récit	711	14	21	0,981	0,971	0,976
Théâtre	403	5	23	0,988	0,946	0,966
Wikipédia	744	6	34	0,992	0,956	0,974

TABLE 3 – Résultats de l'évaluation du tokéniseur du TreeTagger.

Nous avons étudié les erreurs de tokénisation les plus fréquentes. Elles concernent essentiellement des caractères ambigus (*r*, *d*, *z*, *s*) qui doivent ou ne doivent pas être détachés selon le contexte : ainsi *'r* doit être détaché dans *dass'r's* mais pas dans *widd'r* (à noter que dans *dass'r's* il y a deux segmentations à opérer). Ces cas sont difficiles à gérer, compte-tenu des nombreuses variantes graphiques qui peuvent être trouvées (par exemple *us-em* et *us'm*) et de l'apostrophe qui peut être utilisée à la fois pour marquer des élisions en frontière de token et au milieu de mots qui ne doivent pas être découpés (ex : *Z'erscht*, équivalent à *zuerst* en allemand standard). Les deux outils sont de ce point de vue plutôt conservateurs car le nombre de faux négatifs est supérieur au nombre de faux positifs.

4 Tokénisation pour le picard

4.1 Recommandations existantes

Le picard n'existe pas en tant que langue standardisée, ni normée sur le plan de la graphie, cependant, on l'écrit depuis plusieurs siècles déjà et sa présence sur la scène littéraire, sous ses différentes variétés, est importante, ainsi qu'en témoignent de récents succès de librairie et notamment celui des volumes de bandes dessinées en picard d'Astérix et Tintin [1]. Malgré les nombreux débats dans les années 1960-1970 autour de la standardisation et de l'orthographe en picard (centralisation de la langue, question de la variation : uniformisation de la graphie et du lexique), à ce jour, aucune standardisation n'est engagée sur l'ensemble du domaine picard. Plus encore, cette situation est vécue

1. Il n'y a donc pas de contradiction formelle entre la non-standardisation du picard à l'échelle du domaine linguistique picard, la présence de variation et la question de la vitalité notamment (mais pas exclusivement) littéraire du picard. En effet, il n'y a pas ici de vérité ni de lien *stricto sensu* entre standardisation et vitalité sur l'ensemble du domaine picard. Par ailleurs, c'est peut-être ce particularisme de la non-standardisation qui évite de poser un « carcan » trop rigide revendiquant ainsi par la variation une forme de liberté.

aujourd'hui comme une « liberté assumée » contre la standardisation. À ce titre, le picard peut être défini comme une « langue de la liberté » (Martin, 2015). Comme l'écrit Jean-Michel Éloy :

> « [...] en domaine picard, la question de la standardisation n'est pas posée du tout en ce qui concerne les formes de langue. Même si l'on ne parle que de standardisation des procédés graphiques, son importance est diversement appréciée. Les débats sur "l'orthographe du picard", qui furent très vifs dans les années 60 et 70, sont aujourd'hui curieusement éteints – un colloque sur ce thème en 2010 avait d'ailleurs conclu au *statu quo*, et avait eu peu d'écho. » (Eloy, 2014, 10-11)

Nombreux sont les ouvrages en picard, qui mentionnent la non existence de standard et la cooptation d'une « liberté assumée » concernant la graphie. Certains auteurs mentionnent leurs références pour la graphie (Debrie, 1996, 1972; Carton, 1963, 1964, 2001).

4.2 Critères de découpage de mots et tokénisation automatique en picard

Cette variation graphique du picard pose des problèmes pour la tokénisation automatique. Nous nous sommes appuyés sur l'ouvrage *Éche pikar bèl é rade* (Debrie, 1983a), pour proposer des critères du découpage de mots en picard. Les cas les plus problématiques sont les suivants :

- le tiret peut être un séparateur (*Est-ce-què*) mais il peut faire partie de certains mots (par exemple dans certains verbes) : *quandis n'mariye-té* [picard de Belgique] / *kan k i s'marite* [picard de l'Amiénois] (quand elles se marient) ;
- le point peut être utilisé comme séparateur de phrase, entrer dans la composition d'un sigle ou signaler un allongement de la consonne qui conduit à une nasalisation) *I se proumon.ne* [picard du Cambrésis] (il se promène) ;
- l'apostrophe peut avoir plusieurs interprétations possibles : (a) une marque qui oriente par rapport à la prononciation comme dans les exemples *té mérit'roès (tu mérites), f'rais (ferais)*, où une voyelle est supprimée ; (b) une marque de l'élision d'une voyelle *L'aute* (l'autre) ; (c) en fin de mot *Dis-l'*. Il est fréquent d'avoir plusieurs apostrophes dans le même mot avec utilisations différentes *Qu't'os (que tu as), Coreed'l'histoire (encore de l'histoire)* ;
- l'espace. On peut avoir des séquences espace + lettre + espace. Il s'agit du phénomène d'épenthèse, la lettre marque une liaison entre les deux mots : par exemple la lettre z dans *lé z éfans* (les enfants).

La graphie non-standardisée soulève des nombreux problèmes pour le découpage automatique du picard. Outre l'ambiguïté des séparateurs, tel l'apostrophe ou le tiret, l'espace peut apparaître après l'apostrophe et certains pronoms peuvent être agglutinés au mot précédent ou suivant. Ce dernier cas a été traité différemment de l'alsacien. Pour le découpage automatique des mots en picard, nous utilisons, comme pour l'alsacien, des expressions régulières adaptées. D'abord, nous annotons les mots composés incluant le tiret ou l'espace, à l'aide d'un lexique de mots composés construit à partir de plusieurs lexiques picards (Debrie, 1987, 1986, 1985, 1983b, 1981, 1975). Ce lexique contient des locutions prépositionnelles (*a travér dech'*) ou adverbiales (*Tout ein heùt*), des expressions figées (*pi vlaù qu'*). Les mots composés annotés à l'aide du dictionnaire ne seront pas soumis à la procédure de découpage. Ensuite, nous utilisons les séparateurs non-ambigus (!, ;) et nous proposons des règles de découpage spécifiques pour les séparateurs ambigus :

- l'apostrophe est identifiée comme séparateur à l'aide d'une liste de mots outils (déterminants, pronoms, démonstratifs). La présence d'un de ces mots outils avant l'apostrophe, en début ou à la fin du mot, indique qu'on doit découper le mot au niveau de l'apostrophe ;
- le point est considéré comme séparateur de phrases, sauf pour les sigles, les nombres, les marques de nasalisation ;

Graphie initiale	Tokénisation proposée	Phénomène	Source
quand is n'mariye-té (quand elles se marient)	quand␣is n'␣mariye-té	conjonction + pronom, négation + verbe	(Debrie, 1983a)
Est-ce-què	Est␣ce␣què	verbe + pronom + conjonction	(Debrie, 1983a)
I se proumon.ne (il se promène)	I se proumon.ne	pronom + pronom + verbe	(Debrie, 1983a)
O z avon (nous avons)	O z avon	pronom + consonne d'appui + auxiliaire	(Debrie, 1983a)
Té mérit'roès qu'j'el diche à tin père, quand qu'il arvarro (Tu mériterais que je le dise à ton père, quand il arrivera)	Té mérit'roès qu'␣j'␣el diche à tin père␣, quand qu'␣il arvarro	pronom relatif + pronom personnel + pronom, conjonction + pronom	(Debrie, 1983a)
Eze z'éfan i s'abiye (les enfants s'habillent)	Eze z'␣éfan i s'␣abiye	consonne d'appui + nom, pronom personnel + verbe	(Debrie, 1983a)

TABLE 4 – Exemples de découpages en mots proposés pour le picard.

– Le tiret est reconnu comme séparateur dans certains cas particuliers (est-ce-que), sauf pour les mots composés trouvés dans le dictionnaire ou dans certaines formes de verbes ;
– Certains pronoms (*is, i, il*) ou prépositions (*ed'*) agglutinés au mot précédent ou suivant sont découpés en unités indépendantes.

4.3 Évaluation du tokéniseur pour le picard

Nous avons évalué le tokéniseur sur un corpus regroupant des extraits de genres divers (4 191 tokens) : un extrait d'un roman (506 tokens), deux extraits de deux nouvelles (824 tokens), un extrait d'une collection de lettres (452 tokens), un extrait de théâtre (532 tokens), 2 extraits de poésie narrative (635 tokens), 3 extraits de poésie (1 242 tokens). Le genre le plus représenté est la poésie, suivi des nouvelles et de la poésie narrative. En général, ces genres posent des problèmes aux tokéniseurs, à cause de la structure du texte spécifique. Comme pour l'alsacien, nous avons pris en compte le nombre de vrais positifs (VP), de faux positifs (FP) et de faux négatifs (FN) pour calculer la précision, le rappel et la F-mesure. L'évaluation a été faite avec *diff*. Les résultats sont de bonne qualité (tableau 5), en particulier pour la poésie et pour le roman. Les résultats obtenus pour le théâtre et pour la poésie narrative sont les moins précis, ce qui confirme les attentes :

Les erreurs les plus fréquentes relevées dans les résultats du tokéniseur picard sont :
– découpage excessif, dû à la confusion entre un mot outil (réfléchi, conjonction) et le début d'un verbe : *s'roit* sera découpé en *s'*␣*roit* alors qu'il s'agit d'un seul mot ; *qu'meinchi* (commencer) sera découpé alors qu'il s'agit d'un seul mot. Dans certains cas, le point peut également être considéré comme séparateur et le mot se trouve découpé, alors que le découpage ne doit pas se faire ;
– découpage erroné, quand plusieurs découpages sont possibles : *ch'l'* pourra se découper *ch'*␣*l'* ou non selon le contexte. Le mot outil le plus long sera prioritaire.

Genre	VP	FP	FN	Précision	Rappel	F-mesure
Poésie	1 199	19	24	0,984	0,980	0,982
Poésie narrative	563	36	36	0,939	0,939	0,939
Nouvelle	759	36	29	0,955	0,963	0,959
Roman	480	8	18	0,984	0,964	0,974
Lettre	424	12	16	0,972	0,964	0,968
Théâtre	503	13	16	0,974	0,969	0,972

TABLE 5 – Résultats de l'évaluation du tokéniseur pour le picard

Afin de comparer avec un tokéniseur disponible pour le français, notre choix a été l'étiqueteur TreeTagger (Schmid, 1994), le même outil choisi pour la comparaison avec le tokéniseur alsacien. Nous avons comparé notre tokeniseur avec deux configurations différentes de TreeTagger : TreeTagger avec la configuration standard pour le français (TreeTaggerBase) ; TreeTagger adapté pour le picard, utilisant une liste des mots outils en picard finissant par une apostrophe (prépositions, déterminants) et le dictionnaire picard de mots composés (TreeTaggerPicard). Ces deux dernières ressources sont utilisées également par le tokeniseur picard.

Configuration	Genre	VP	FP	FN	Précision	Rappel	F-mesure
TreeTaggerBase	Poésie	1 014	217	202	0,824	0,834	0,829
	Poésie narrative	486	62	119	0,887	0,803	0,843
	Nouvelle	685	68	80	0,910	0,895	0,903
	Roman	426	24	66	0,947	0,866	0,904
	Lettre	415	27	33	0,939	0,926	0,933
	Théâtre	510	15	11	0,971	0,979	0,975
TreeTaggerPicard	Poésie	1 031	196	190	0,840	0,844	0,842
	Poésie narrative	545	58	62	0,904	0,898	0,901
	Nouvelle	714	54	64	0,930	0,918	0,924
	Roman	455	5	48	0,989	0,905	0,945
	Lettre	412	28	35	0,936	0,922	0,929
	Théâtre	515	12	9	0,977	0,983	0,980

TABLE 6 – Résultats de l'évaluation du tokéniseur du TreeTagger français (TreeTaggerBase) et du TreeTagger utilisant les ressources spécifiques au picard (TreeTaggerPicard)

Le tokéniseur picard obtient des meilleurs résultats par rapport à TreeTagger (sauf pour le théâtre). Ces résultats s'expliquent par la mise en place des règles et des ressources adaptées au picard. TreeTagger se distingue du tokéniseur picard par la différence de traitement de l'apostrophe (qui est souvent considérée comme token séparé). TreeTaggerBase a systématiquement obtenu des performances plus faibles que TreeTaggerPicard. Malgré les ressources ajoutées, les résultats du TreeTaggerPicard restent inférieurs aux performances du tokéniseur du picard, à l'exception du théâtre où les tendances sont inversées (F-mesure de 0,980 contre 0,972 pour le tokéniseur picard) (voir tableau 6).

5 Conclusion et perspectives

Nous avons présenté les problèmes liés à la mise en place des outils de tokénisation pour l'alsacien et le picard, deux langues régionales dont l'orthographe est peu standardisée. Nous avons traité certains séparateurs ambigus (tels le point ou l'apostrophe), le cas des mots agglutinés et de l'épenthèse. Le découpage automatique rencontre des difficultés similaires pour les deux langues : plusieurs découpages possibles, plusieurs interprétations pour les séparateurs, ou les phénomènes d'épenthèse. Les erreurs les plus fréquentes sont liées aux séparateurs ambigus. A l'avenir, les outils de tokénisation seront évalués sur des corpus de plus grande taille, incluant d'autres genres (conte, dialogue, etc.) et les règles de découpage seront améliorées. Ils seront utilisés pour le développement des outils de traitement automatique de l'alsacien et du picard (étiqueteur, lemmatiseur).

Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

Références

BENZITOUN C., FORT K. & SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, p. 99–112, Grenoble, France.

CARTON F. (1963). Essai d'adaptation de l'orthographe Feller au picard moderne. *Nos patois du Nord*, **8**(supplément), 134–139.

CARTON F. (1964). L'adaptation de l'orthographe Feller au picard moderne. *Nos patois du Nord*.

CARTON F. (2001). Orthographe picarde Feller-Carton. *Linguistique picarde*.

DEBRIE R. (1972). *Propos sur l'orthographe*. Amiens : Archives départementales de la Somme.

DEBRIE R. (1975). *Lexique picard des parlers ouest-amiénois*. Centre d'Études Picardes.

DEBRIE R. (1981). *Lexique picard du Vimeu*. Centre d'Études Picardes.

DEBRIE R. (1983a). *Eche pikar bèl é rade*. Ed.-disques Omnivox.

DEBRIE R. (1983b). *Lexique picard des parlers est-amiénois*. Centre d'Études Picardes.

DEBRIE R. (1985). *Lexique picard du Ponthieu*. Centre d'Études Picardes.

DEBRIE R. (1986). *Lexique picard des parlers du Santerre*. Centre d'Études Picardes.

DEBRIE R. (1987). *Lexique picard du Vermandois*. Centre d'Études Picardes.

DEBRIE R. (1996). Essai d'orthographe picarde. *Le Courrier Picard*.

J.-M. ELOY, Ed. (2014). *Standardisation et vitalité des langues de France*, volume 9 of *Carnets d'Ateliers de Sociolinguistique*. Paris : L'Harmattan.

GREFENSTETTE G. (1998). *Re : Corpora : Sentence splitting*. Corpora List http ://torvald.aksis.uib.no/corpora/1998-4/0035.html.

GREFENSTETTE G. & TAPANAINEN P. (1994). What is a word, What is a sentence ? Problems of Tokenization. In *3rd International Conference on Computational Lexicography (COMPLEX'94)*, Budapest, Hungary.

JUNG E. (1983). *Grammaire de l'alsacien, dialecte de Strasbourg avec indications historiques*. Strasbourg, France : Oberlin.

JURISH B. & WÜRZNER K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, **28**(2), 61–83.

KECK B. & DAUL L. (2010). *L'alsacien pour les nuls*. Pour les nuls (Éd. de poche), ISSN 1625-0486. Paris, France : First éd.

LAARMANN-QUANTE R. & DIPPER S. (2016). An Annotation Scheme for the Comparison of Different Genres of Social Media with a Focus on Normalization. In *Proceedings of the LREC Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*.

MARTIN F. (2015). *Espaces et lieux de la langue au XXIe siècle en Picardie. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires – enquête en Picardie*. Thèse de doctorat, Université de Picardie Jules Verne, Amiens.

POINTAL L. (2004). Tree Tagger Wrapper. [en ligne ; accédé le 4 avril 2016] `https://perso.limsi.fr/pointal/dev:treetaggerwrapper`.

RABARY C. T., LAVERGNE T. & NÉVÉOL A. (2015). Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France.

REDSLOB R. (1907). *D'r Schlitterhannes. Elsaessisches Bauerndrama in zwei Akten*. Strassburg, 1907.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

SONNENDRÜCKER P. & KAUSS A. (1998). *Kochersberg : récits en dialecte avec version française*. Strasbourg, France : Bf.

STOSKOPF G. (1906). *D'r Hoflieferant. Elsaessische Komœdie in 3 Aufzuegen*. Strassburg, 1906.

WEBSTER J. J. & KIT C. (1992). Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on Computational linguistics-Volume 4*, p. 1106–1110.

WIKIPEDIA (2015). Elsässisches Museum (Straßburg) — Alemannische Wikipedia. [En ligne, accédé le 01/06/2017, Page Version ID : 654412].

WIKIPEDIA (2017). Johannes Mentelin — Alemannische Wikipedia. [En ligne, accédé le 01/06/2017, Page Version ID : 754490].

WRENN J. O., STETSON P. D. & JOHNSON S. B. (2007). An unsupervised machine learning approach to segmentation of clinician-entered free text. In *AMIA Annual Symposium Proceedings*, volume 2007 : American Medical Informatics Association.

ZEIDLER E. & CRÉVENAT-WERNER D. (2008). *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. Colmar : J. Do Bentzinger.