Vers un lexique ouvert des formes fléchies de l'alsacien : génération de flexions pour les verbes

Lucie Steiblé Delphine Bernhard
LiLPa - EA 1339, Université de Strasbourg
{lucie.steible,dbernhard}@unistra.fr

RÉSUMÉ

Cet article présente les méthodes mises en œuvre et les résultats obtenus pour la création d'un lexique de formes fléchies de l'alsacien. Les dialectes d'Alsace font partie des langues peu dotées : rares sont les outils et ressources informatisées les concernant. Plusieurs difficultés doivent être prises en compte afin de générer des ressources pour ces langues, généralement liées à la variabilité en l'absence de norme graphique, et au manque de formes fléchies dans les quelques ressources existantes. Nous avons pour ce faire utilisé plusieurs outils permettant la génération automatique de variantes graphiques et la création de formes fléchies (graphes morphologiques et de flexion d'Unitex). Les résultats en termes de couverture des formes rencontrées dans des textes ont permis l'évaluation de la méthode.

ABSTRACT

Towards an Open Lexicon of Inflected Word Forms for Alsatian: Generation of Verbal Inflection.

This article presents the methods used and results obtained for the creation of a lexicon of inflected word forms for Alsatian. The dialects from Alsace are low resourced languages: few computational tools and resources are available. When creating such resources, many difficulties have to be addressed, usually related to variability issues in the absence of standard spelling and the lack of inflected forms in the few existing resources. We used several tools for the automatic generation of graphical variants and the creation of inflected forms (morphological and inflection graphs of Unitex). The method was evaluated in terms of coverage of the verbal forms encountered in texts.

MOTS-CLÉS : Lexique, flexion, verbes, variabilité graphique, Unitex, alsacien.

KEYWORDS: Lexicon, inflection, verbs, graphic variants, Unitex, Alsatian.

1 Introduction

L'outillage informatique des langues peu dotées se heurte souvent au problème de la grande variabilité des formes rencontrées dans les textes, en l'absence de norme graphique établie. Ainsi, la construction de lexiques de formes fléchies doit prendre en compte à la fois les variantes flexionnelles et les variantes graphiques des lemmes. Nos travaux concernent plus particulièrement les dialectes d'Alsace, pour lesquels il n'existe pas encore de lexique de formes fléchies. Les ressources numériques existantes listent les mots uniquement sous leur forme canonique, avec souvent plusieurs variantes graphiques par lemme. Cet article décrit la méthode employée pour constituer un lexique limité dans un premier temps aux verbes, qui ont les paradigmes flexionnels les plus riches par rapport aux autres catégories grammaticales, même si le système de flexion reste relativement simple (voir Section 3).

La génération des formes fléchies utilise les graphes de flexion du système Unitex[1] (Paumier, 2016), couplés à la génération automatique de variantes graphiques et la détection des formes préfixées de verbes (Section 4). La couverture de la ressource est évaluée par rapport à des textes préalablement annotés morphosyntaxiquement (Section 5).

2 État de l'art

Il existe plusieurs lexiques de lemmes et formes fléchies pour le français (Dictionnaires électroniques du LADL (Courtois, 1990), Morphalou (Romary *et al.*, 2004), Lexique3 (New, 2006), VfrLPL (Rauzy & Blache, 2007), Lefff (Sagot, 2010), Glàff (Sajous *et al.*, 2013)) et l'allemand (DeLex (Sagot, 2014), Zmorge (Sennrich & Kunz, 2014), etc.), pour ne citer que ces deux langues. Ces ressources diffèrent par leur couverture (pour le français, voir l'étude de Sajous *et al.* (2013)), leur format (XML, csv), leurs contenus (présence ou non d'informations phonologiques ou de fréquence, entre autres) et leur mode de construction. De manière à faciliter la maintenance et l'évolution des lexiques, les formes fléchies sont souvent générées automatiquement à partir du lexique de lemmes. La forme canonique est transformée de manière à produire en sortie la forme fléchie associée à des informations flexionnelles (mode, temps, personne pour les verbes). La définition des transformations à opérer nécessite un travail linguistique préalable, afin de modéliser les classes de formes qui partagent les mêmes opérations de flexion : dans le cas des verbes, il s'agira de modéliser les conjugaisons. Le nombre de classes est dépendant de la complexité et des irrégularités du système flexionnel. Ce principe de génération automatique de formes fléchies est notamment employé pour la version extensionnelle du Le*fff* (Sagot, 2010), reposant sur le système Alexina.

Les travaux les plus récents s'intéressent à la construction automatique de lexiques à partir de ressources collaboratives du type wiki, afin de tirer parti de l'actualité des informations que l'on peut y trouver. Glàff par exemple est construit à partir du Wiktionnaire français, par extraction d'informations à partir des articles, y compris les formes fléchies et les informations morphosyntaxiques (Sajous *et al.*, 2013). Pour l'allemand, Sennrich & Kunz (2014) extraient un lexique compatible avec une grammaire morphologique à états finis : le défi principal consiste à prédire la classe de flexion existante dans la grammaire. Une méthode proche est employée pour DeLex (Sagot, 2014) : extraction des entrées du Wiktionary allemand, construction automatique de classes flexionnelles partielles et définition manuelle des classes finales. Dans nos travaux sur les dialectes alsaciens, nous reprenons le principe de génération automatique de formes fléchies à partir des lemmes groupés en classes de flexions.

3 Spécificités de l'alsacien

3.1 Conventions orthographiques

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. Par exemple, le verbe «jouer» est présent sous quatre formes différentes dans notre lexique : *spìela, spiele, speele, schpeela*. Cette variabilité est à la fois liée à des différences de prononciation, mais également aux choix de transcription graphique des auteurs.

[1] http://www-igm.univ-mlv.fr/~unitex/

3.2 Temps verbaux en alsacien

Le système temporel et aspectuel de l'alsacien est différent de celui du français et de l'allemand. Il repose sur seulement trois formes : une simple, une composée et une surcomposée qui permettent d'exprimer les temps et les aspects (Kleiber & Matterer, 1977). Par exemple, le passé est composé, et le futur n'est généralement pas exprimé par un marquage morphologique mais par l'utilisation d'éléments extra-verbaux. Ainsi, trois temps sont principalement utilisés en alsacien : le présent, le passé composé et le plus-que-parfait, le prétérit n'étant plus en usage depuis la fin du XVI^{e} siècle (voir Table 1). Les modes sont également au nombre de trois : l'indicatif, l'impératif et le subjonctif. La morphologie verbale indique également la personne, mais toutes les formes du pluriel sont identiques à la forme de l'infinitif. Cette relative simplicité facilite la modélisation pour la flexion automatique, puisque les formes à obtenir sont en nombre réduit et couvrent la plupart des usages, en dehors de quelques exceptions qui peuvent être implémentées manuellement (comme les quelques formes synthétiques qui existent pour de rares verbes au subjonctif ou à l'impératif).

Présent	Ich **sing**	Je chante
Futur	Ich **sing** morne	Je chanterai (demain)
Passé composé	Ich **hàb** a Waawe **kauft**	J'ai acheté une voiture
Plus-que-parfait	Ich **hàb** a Waawe **kauft ghett**	J'avais acheté une voiture

Table 1: Formes verbales de l'alsacien

4 Génération des formes verbales fléchies

4.1 Sources lexicales

Les lemmes utilisés pour la génération sont des formes attestées dans les lexiques et dictionnaires suivants :

1. Le lexique de l'Association Culture et Patrimoine d'Alsace[2] (3 409 lemmes verbaux)
2. Les lexiques thématiques de l'Office pour la Langue et la Culture d'Alsace[3] (783 lemmes verbaux)
3. Le Dictionnaire Comparatif Multilingue (Adolf, 2006) (609 lemmes verbaux)

A cela s'ajoutent des verbes cités dans diverses grammaires (Huck *et al.*, 1999; Jenny & Richert, 1984; Jung, 1983; Keck *et al.*, 2010; Nisslé, 2013).

Chaque entrée du lexique des lemmes se compose des informations suivantes :

- le lemme
- le code morphologique, qui fait référence à la classe de conjugaison (voir Figure1a pour le code `Vpartmit`, classe des verbes à particule *«mit»*)

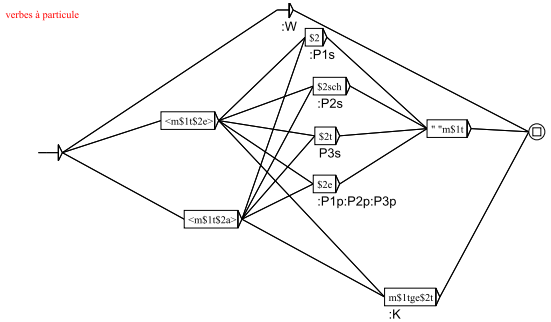
[2]http://culture.alsace.pagesperso-orange.fr/
[3]https://www.olcalsace.org/fr/lexiques

- divers champs supplémentaires : les sources (par exemple `acpa`), les indications géographiques (`MA`, Moyenne Alsace), et les traductions en français. Les entrées ressemblent alors à ce modèle : `metmàcha,Vpartmit+acpa+MA+fr#participer`.

21 codes morphologiques sont nécessaires en alsacien, afin de respecter les spécificités des verbes lors de la conjugaison. La catégorie `V1` regroupe les verbes dits réguliers. Lors de la conjugaison de ces verbes, un suffixe ***sch*** est ajouté afin de générer la forme de la deuxième personne du singulier : par exemple *redde* (parler) → *du redd**sch*** (tu parles). Les verbes qui finissent en ***sche*** comme *ewerrasche* (surprendre) sont donc dits irréguliers, puisque la conjugaison de la deuxième personne du singulier ne saurait suivre la règle habituelle : *du ewerasch**sch*** est incorrect selon toutes les grammaires. Il faut donc considérer toutes les catégories de verbes en fonction de leur comportement morphologique. Chaque catégorie peut alors recevoir un code morphologique afin de procéder à l'étape suivante, la flexion automatique à proprement parler.

4.2 Flexion automatique

Unitex permet de fléchir automatiquement des formes canoniques à l'aide de graphes de flexion. Cet outil présente l'avantage de proposer une interface graphique et une documentation fournie, facilitant ainsi le travail de mise au point des graphes. La forme canonique est exploitée en entrée du graphe, avant d'être transformée par ajout des terminaisons. Chaque forme ainsi obtenue est étiquetée avec son code DELAF. L'alsacien ayant un système verbal relativement simple, le nombre de formes nécessaires est faible : l'infinitif, qui ne subit aucune modification (`:W`) celles du présent à toutes les personnes (`:P1s`, `:P2s`, etc.) et celle du participe passé (`:K`).



(a) Graphe de flexion pour les verbes à particule «*met*», «*mit*» ou «*mìt*»

```
metmàcha,metmàcha.Vpartmit+acpa+inc+fr#participer:W
màch met,metmàcha.Vpartmit+acpa+inc+fr#participer:P1s
màchsch met,metmàcha.Vpartmit+acpa+inc+fr#participer:P2s
màcht met,metmàcha.Vpartmit+acpa+inc+fr#participer:P3s
màche met,metmàcha.Vpartmit+acpa+inc+fr#participer:P3p
màche met,metmàcha.Vpartmit+acpa+inc+fr#participer:P2p
màche met,metmàcha.Vpartmit+acpa+inc+fr#participer:P1p
metgemàcht,metmàcha.Vpartmit+acpa+inc+fr#participer:K
```

(b) Formes fléchies du lemme «*metmàcha*» après sa flexion par le graphe

Figure 1: Exemple de verbe à particule dans Unitex

La Figure 1a présente un graphe permettant d'obtenir les formes du présent ainsi que le participe passé des verbes qui, à l'infinitif, commencent par la particule séparable «*met*» (avec). Cette particule doit être séparée de la base pour la conjugaison : *metmàcha* (littéralement «faire avec», participer) doit devenir pour la première personne du singulier *Ech màch met* (Je participe). Il est possible de conserver la base du verbe dans une variable, ici $2, afin de lui adjoindre les terminaisons tout en déportant la particule. La variable $1 quant à elle permet de gérer les formes différentes de «*met*» que l'on pourrait trouver : «*mit*» ou «*mìt*» par exemple. Tout caractère entre le «m» et le «t» sera stocké dans la variable de la particule.

Les graphes permettent donc de générer des formes fléchies et étiquetées selon leur nature : infinitive, au participe, ou accordées selon les différentes personnes du présent. A la suite de cette étape, le lexique de 4 801 lemmes verbaux a permis de générer 26 342 formes fléchies, soit une multiplication par 5 des formes connues.

4.3 Génération de variantes graphiques

Nous générons également des variantes graphiques supplémentaires pour les formes fléchies. La génération des variantes graphiques s'inspire du mode de fonctionnement de Varialog (Lay, 2012), développé à l'origine pour gérer la variation graphique dans des textes français de la Renaissance. La méthode consiste à définir des règles contextualisées de génération de variantes, de manière à générer des formes qui pourraient se trouver dans les textes et améliorer ainsi la couverture du lexique. Les règles de variation ont été identifiées automatiquement à partir de l'alignement de formes trouvées dans divers lexiques alsacien-français et allemand-français (Bernhard & Steiblé, 2015). Elles se présentent sous la forme suivante : `^un = ùn` (la chaîne `un` située en début de mot sera remplacée par `ùn`), `iaga$ = ije` (la chaîne `iaga` située en fin de mot sera remplacée par `ije`). Afin d'éviter la génération d'un trop grand nombre de formes peu plausibles, les remplacements sont contextualisés par un caractère à gauche et à droite, début et fin de mots compris. Au total, 1 438 règles de générations de variantes de fréquence supérieure à 5 ont été utilisées (ces règles peuvent s'appliquer dans les deux sens). Après génération des variantes, le lexique de verbes conjugués comprend 6 137 049 formes, soit environ 230 fois plus que le lexique fléchi initial. Il faut noter ici qu'il s'agit d'une sur-génération : toutes les formes générées ne sont pas attestées, ni même vraisemblables.

4.4 Détection des préfixes et bases

Les bases verbales en alsacien sont susceptibles d'être préfixées par des particules séparables ou inséparables. Le problème est alors que la base ne sera pas reconnue, bien que présente dans le lexique de lemme. Il est possible de séparer les préfixes afin de permettre la détection de la base, en listant les préfixes possibles dans un graphe morphologique d'Unitex (Paumier, 2016, section 6.4.3 Dictionnaires du mode morphologique). Grâce à cette opération, des formes absentes du lexique, telles que *inkoche* (faire réduire) ou *àbkoche* (bouillir), peuvent être reconnues comme variantes de *koche* (cuire), qui est dans le lexique. Le graphe morphologique permet la reconnaissance de 38 préfixes usuels.

5 Analyse de la couverture de la ressource

La couverture de la ressource a été évaluée sur un corpus de 4 textes annotés manuellement avec les catégories grammaticales. La ressource a été appliquée au corpus de trois manières différentes :

1. Lexique initial : le lexique de verbes fléchis automatiquement a été appliqué directement à l'aide d'Unitex ;
2. Lexique expansé : des variantes graphiques des formes conjuguées ont été générées automatiquement et le lexique expansé a été appliqué à l'aide d'Unitex.
3. Lexique expansé + préfixes : les formes préfixées ont été détectées à l'aide d'un graphe morphologique.

Les textes testés comportent 179 formes verbales au total. La Table 2 présente les résultats obtenus.

Formes verbales reconnues			**Formes verbales non reconnues**		
Formes fléchies seulement	68	= 38%	Formes différentes	38	= 22%
+ Formes expansées	19	→ 49%	Formes absentes des lexiques	36	= 20%
+ Formes préfixées	18	→ 58%			
Total des formes reconnues	**105**	**= 58%**	Total des formes non reconnues	**74**	= 42%

Table 2: Couverture de la ressource (résultats absolus à gauche, en pourcentage à droite).

On constate la rentabilité de l'implémentation de formes expansées, qui permet la reconnaissance de *schniide* ou encore *pfeffere* sur la base de *schnide* (couper) et *pfafere* (poivrer) qui étaient les formes présentes dans le lexique originel. Cette opération permet de reconnaître 19 formes supplémentaires. L'utilisation du graphe morphologique pour les formes préfixées est également efficiente, puisqu'elle permet la reconnaissance de *màcht* (faire, 3e personne du singulier) dans *ufgmàcht* ou encore de *fàhrt* (partir, 3e personne du singulier) dans *erfàhrt*. La reconnaissance de ces formes permet un gain de 18 lemmes supplémentaires. La couverture atteint donc 105 formes sur 179, et permet de constater le gain sensible d'efficacité permis par les étapes d'expansion et de reconnaissance des préfixes.

Les formes non reconnues sont, au total, au nombre de 74. La non reconnaissance est due, dans la moitié des cas, à l'absence du lemme dans le lexique, et pour l'autre moitié à un éloignement trop grand entre la forme lexique et celle rencontrée dans les textes. C'est le cas pour *bstraie* (saupoudrer), *lon* (laisser), *höre* (entendre) qui n'ont malheureusement pas pu être rapprochées des formes des lexiques *beschtraije*, *losse* / *loh* / *lo* et *heere* / *hüre*. Il existe aussi 8 cas de faux positifs (4%), lorsqu'une forme a bien été reconnue comme verbe, mais a été rapprochée d'un lemme inexact. On peut observer la forme *gsajt* (dire, au participe passé) qui a été reconnue en tant que forme issue du lemme *saje* (scier) au lieu du lemme *sage* (dire). Ces formes n'ont pas été comptée comme correctes, bien que la catégorie grammaticale ait bien été reconnue. Seules 2 formes (1%) ont été rapprochées à tort d'un lemme suite à la sur-génération graphique, ce qui semble raisonnable comparé aux 10% de reconnaissances correctes acquises grâce à cette sur-génération.

6 Conclusion et perspectives

La couverture de 58% des formes peut sembler faible, mais eu égard à la complexité d'analyse des langues non normées, ces résultats sont encourageants. Les performances pourraient en effet être améliorées en suivant plusieurs axes de travail. Le principal problème rencontré pour la reconnaissance des formes est la variabilité graphique des dialectes d'Alsace : 38 formes n'ont pas été rapprochées de lemmes pourtant présents dans le lexique. Une gestion plus fine des règles d'expansion du lexique permettrait d'améliorer cet aspect. Par exemple, il sera utile de ne prendre en compte que les remplacements les plus fréquents au sein des lemmes verbaux et non au sein de l'ensemble du lexique. L'autre problème majeur concerne l'absence complète de certains verbes dans le lexique. Ces formes seront ajoutées au fur et à mesure à partir d'études de corpus. Enfin, il sera également utile de scinder certaines formes préfixées du lexique, afin d'ajouter la forme de la base dans le lexique, permettant d'obtenir de plus nombreuses variantes attestées du même lemme.

Remerciements

Nous remercions Pascale Erhart pour sa relecture et les auteurs de lexiques et dictionnaires pour nous avoir donné accès à leurs ressources. Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01).

References

ADOLF P. (2006). *Dictionnaire comparatif multilingue: français-allemand-alsacien-anglais*. Strasbourg, France: Midgard.

BERNHARD D. & STEIBLÉ L. (2015). Quand l'oral se fait entendre à l'écrit : alignement de lexiques en l'absence de normalisation graphique. In *Actes de l'atelier sur le Traitement Automatique des Langues Régionales de France et d'Europe*, Caen, France.

COURTOIS B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, (87), 11–22.

HUCK D., LAUGEL A. & LAUGNER M. (1999). *L'élève dialectophone en Alsace et ses langues : L'enseignement de l'allemand aux enfants dialectophones à l'école primaire*. Strasbourg: Oberlin.

JENNY A. & RICHERT D. (1984). *Précis pratique de grammaire alsacienne en référence principalement au parler de Strasbourg*. Strasbourg: ISTRA.

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

KECK B., DAUL L. & KRETZ P. (2010). *L' alsacien pour les nuls*. Pour les nuls. Paris.

KLEIBER G. & MATTERER P. (1977). *Une première approche du verbe en alsacien*. Colmar, France: C.D.D.P.

LAY M.-H. (2012). VariaLog: how to locate words in a French Renaissance Virtual Library. In *Proceedings of Digital Humanities 2012*, Hamburg.

NEW B. (2006). Lexique 3: Une nouvelle base de données lexicales. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*.

NISSLÉ A. (2013). *D'Lehrschtuwa, la grammaire alsacienne : des règles expliquées, des exemples simples, des conseils pour s'exprimer*. Mulhouse: Association Culture et Patrimoine d'Alsace Editions JdM.

PAUMIER S. (2016). *Unitex 3.1 user manual*.

RAUZY S. & BLACHE P. (2007). Un lexique syntaxique des verbes du français: VfrLPL. *Rapport de recherche RAU*, **3055**.

ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop On Enhancing And Using Electronic Dictionaries at the 20th International Conference on Computational Linguistics - COLING 2004*, Genève/Switzerland.

SAGOT B. (2010). The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

SAGOT B. (2014). DeLex, a Freely-available, Large-scale and Linguistically Grounded Morphological Lexicon for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). Glàff, un gros lexique à tout faire du français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298.

SENNRICH R. & KUNZ B. (2014). Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

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