SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif

Franck Sajous, Nabil Hathout et Basilio Calderone
CLLE-ERSS (CNRS & Université de Toulouse-Le Mirail)
{franck.sajous,nabil.hathout,basilio.calderone}@univ-tlse2.fr

1 Introduction

Wiktionnaire[1] est l'édition française de Wiktionary, le dictionnaire libre multilingue accessible en ligne. Satellite de Wikipédia, dont il constitue le *« compagnon lexical »*, le projet dictionnairique reste dans l'ombre de l'encyclopédie. Fondé comme elle sur le principe du wiki, il peut être alimenté et modifié par tout internaute, avec publication immédiate. Si la ressource encyclopédique a été abondamment utilisée dans certaines disciplines, le dictionnaire collaboratif semble avoir reçu moins d'attention de la part de la communauté scientifique. Ce moindre intérêt pourrait être le fruit d'une méconnaissance ou d'un rejet *a priori* de l'amateurisme que l'on associe volontiers aux contributions effectuées par des naïfs.
Nous présentons dans cet article quelques caractéristiques du Wiktionnaire et des réalisations issues de cette ressource. Ce travail illustre différentes possibilités offertes par ce dictionnaire singulier, comment il est possible d'exploiter ses spécificités, dans quelle mesure, et pour quel usage. La question centrale abordée dans cet article est la possibilité d'utiliser le Wiktionnaire comme point de départ pour la constitution d'un lexique électronique pour des domaines comme le traitement automatique des langues (TAL) d'une part, et en complément de ressources destinées à des études linguistiques ciblées d'autre part.

L'article s'organise comme suit : nous dressons dans la section 2 un panorama des ressources lexicales électroniques existantes et expliquons pourquoi une ressource encore très perfectible comme le Wiktionnaire a retenu notre attention. Nous questionnons ensuite le statut des ressources collaboratives et la légitimité de leur utilisation dans des travaux de recherche, puis donnons des éléments de description du Wiktionnaire (section 3). Nous montrons enfin comment nous en avons extrait plusieurs lexiques flexionnels et phonologiques à large couverture pour le TAL, la linguistique outillée et la psycholinguistique (section 4).

2 Ressources lexicales : une situation insatisfaisante

Plusieurs ressources lexicales pour le français commencent à être disponibles, même s'il reste beaucoup à faire pour nous rapprocher de la situation de l'anglais, tant en volume qu'en qualité. Le constat est similaire pour les outils généralistes comme les analyseurs morphosyntaxiques, syntaxiques, morphologiques et les outils de phonétisation, qui dépendent directement de ces ressources. Depuis la fin des années 1990, quelques ressources destinées au traitement automatique du français sont distribuées gratuitement : le lexique de l'ABU[2] date de 1999, la première version de Lefff (Clément et al., 2004) de 2003 et Morphalou (Romary et al., 2004) de 2004. Auparavant, seules des ressources payantes, principalement distribuées par ELRA, étaient disponibles. ABU, le plus ancien des lexiques morphosyntaxiques distribué librement sur le Web, comporte environ 300 000 formes et 60 000 lemmes (ou formes de citation). Lefff et Morphalou comptent quant à eux respectivement 500 000 et 525 000 entrées. Lefff fournit également une description

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801216

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

des cadres de sous-catégorisation des lexèmes.

La constitution de lexiques pour le TAL et pour l'étude outillée du français trouve son origine dans les travaux menés au LADL autour de Maurice Gross (Courtois, 1990; Silberztein, 1990). Ces lexiques étaient destinés à l'exploration de corpus et l'annotation lexicale. Les lexiques morphosyntaxiques du français fournissent tous un ensemble d'informations communes : la forme orthographique du mot, son lemme, la partie du discours et des propriétés morphosyntaxiques (traits grammaticaux et flexionnels). Ces ressources, d'abord utilisées pour le TAL, le sont également pour la description linguistique, notamment en morphologie (Hathout et al., 2009a).

À côté des ressources lexicales développées par les linguistes informaticiens et les lexicographes[3], on trouve Lexique (New, 2006), qui s'inscrit dans la lignée de Brulex (Content et al., 1990), développée à la fin des années 1980. Comme Brulex, Lexique a été créée par et pour les psycholinguistes. Ces ressources se distinguent des lexiques morphosyntaxiques généralistes par la plus grande richesse des informations fournies : outre la morphosyntaxe, leur description lexicale comporte une transcription phonémique[4], une segmentation en syllabes, des informations sur les homophones, les homographes, les voisins phonologiques et orthographiques, la fréquence des formes dans des corpus écrits, etc. En contrepartie, un grand nombre de formes fléchies en sont absentes (seules les formes les plus usuelles y sont décrites) et les fréquences fournies, calculées à partir des graphies, ne tiennent pas compte des attributs morphosyntaxiques. Lexique et Brulex étaient, jusqu'à récemment, les seules ressources gratuites fournissant des transcriptions phonémiques et un découpage syllabique. Il existe d'autres ressources avec une couverture plus complète, contenant ces mêmes informations, qui, bien qu'issues de laboratoires de recherche publique, sont payantes et, surtout, sous licences non libres. L'une des plus anciennes et la plus connue est BDLex (Pérennou et de Calmès, 1987), dont la taille est similaire à celle de Lefff. Citons également ILPho (Boula De Mareuil et al., 2000), plus récente, créée en complétant les entrées du lexique morphosyntaxique Multext (Ide et Véronis, 1994) par des transcriptions phonémiques. Enfin, GlobalPhone (Schultz et al., 2013) est une base de données comprenant des dictionnaires de prononciation, dont la première version date de 2002. Ces dictionnaires ont été produits automatiquement par des outils entraînés sur des transcriptions d'enregistrements de locuteurs et les textes correspondants, puis corrigés manuellement. Aujourd'hui disponible pour 20 langues, cette ressource, comme ILPho et BDLex, est vendue par ELRA[5].
Enfin, une tendance a vu ces dernières années l'apparition de ressources volumineuses telles que WOLF (Sagot et Fišer, 2008) ou BabelNet (Navigli et Ponzetto, 2010) produites automatiquement et, du fait de leur taille, non validables manuellement. Ces ressources résultent souvent de l'agrégation et de la traduction automatique d'autres ressources et d'application de règles d'inférence.

Ainsi, il n'existe pas à l'heure actuelle de lexique électronique de qualité pour le français, qui présente les caractéristiques attendues d'une telle ressource : 1) une large couverture 2) des transcriptions phonémiques 3) une licence libre 4) une mise à jour régulière. Dans ce contexte, il nous a paru intéressant d'étudier les potentiels du Wiktionnaire, qui pourraient satisfaire l'ensemble de ces exigences.

3 Le Wiktionnaire

3.1 La *« sagesse des foules »* : quelle légitimité ?

Un article de Giles (2005), paru dans *Nature*, avait soulevé une polémique en prétendant que la qualité des articles de Wikipédia était comparable à celle des articles de l'encyclopédie Britannica. Cette dernière a réfuté la comparaison en mettant en cause les critères d'évaluation utilisés (Encyclopaedia Britannica, 2006). À l'opposé de Giles, Pierre Assouline (2007) qualifiait Wikipédia, dans *L'Histoire*, d'*« erreur à haut débit »*. Ces antagonismes, certainement tous deux excessifs, interrogent sur le statut des contenus *« approvisionnés par les foules*[6] *»*.

Les premiers travaux utilisant Wiktionary ont été menés par Zesch et al. (2008) pour effectuer des calculs de similarité sémantique. Se donnant cette même tâche comme critère d'évaluation, Zesch et Gurevych (2010)

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

comparent la qualité des ressources fondées sur « la sagesse des linguistes » et celles construites collaborativement « par les foules ». Ils concluent que ces dernières ne sont pas meilleures que celles construites par les experts, mais sont sérieusement compétitives et les dépassent en termes de couverture. L'étude portait cependant sur l'utilisation de données dérivées de Wiktionary et non sur son contenu primaire. Seule une étude systématique d'un nombre suffisamment important d'articles nous permettrait de juger de la qualité intrinsèque du Wiktionnaire en tant que dictionnaire. Une telle étude reste à faire, mais gageons qu'elle conclurait à une qualité moindre du Wiktionnaire par rapport aux dictionnaires choisis comme référence. Pour autant, nous ne pensons pas qu'il faille rejeter en bloc le Wiktionnaire et s'en interdire l'usage. La défiance à l'égard des ressources collaboratives n'est pas sans rappeler celle avec laquelle a été initialement accueillie la proposition de recourir au Web, notamment en linguistique de corpus. La problématique du *« Web as Corpus »* a bénéficié de réflexions comme celles de Kilgarriff et Grefenstette (2003) qui, à la question de savoir si le Web est un corpus, préfèrent se poser la suivante : *« le corpus X est-il pertinent pour la tâche Y ? »* (le Web est un corpus… représentatif de lui-même). Depuis, l'utilisation du Web a permis de mener des études qui n'auraient pas pu être réalisées autrement. En morphologie extensive par exemple, nous avons montré dans une étude portant sur les noms déverbaux suffixés en *-ion*, *-age* ou *-ment*, qu'il est nécessaire de parcourir en moyenne 1000 pages web pour trouver un dérivé valide inconnu (Hathout et al., 2009b). Seul le Web est susceptible de fournir la quantité de textes requis. De la même manière, le Wiktionnaire ne peut prétendre au statut de dictionnaire au même titre que les ouvrages de référence, mais offre des possibilités inédites. Nous nous rangeons au côté de Penta (2011) qui rejette les sarcasmes exagérés ciblant Wiktionary ou Urban Dictionary[7] et répond aux détracteurs de ces ressources en montrant qu'elles sont au moins complémentaires des dictionnaires traditionnels. Un atout des dictionnaires en ligne est l'absence de contrainte de taille. Cette liberté leur permet de contenir des variations diatopiques et diastratiques que l'on ne trouve pas nécessairement dans les dictionnaires imprimés. Ils sont par ailleurs plus réactifs à la néologie : la frilosité avec laquelle les créations lexicales font leur apparition dans les lexiques traditionnels n'a pas de raison d'être dans le cas des dictionnaires collaboratifs. Dal et Namer (2012), questionnant la notion d'existant du lexème, mentionnent la tension pour les morphologues entre l'étude d'un lexique institutionnalisé issu des dictionnaires et l'utilisation de données écologiques, issues par exemple du Web. Elles donnent des exemples de mots très fréquents (nombre d'occurrences sur le Web à l'appui), qui, absents des dictionnaires, pourraient échapper à la description morphologique (e.g. *américanité, bien-penseance, contre(-)productif, doublonner, européanité, fiabiliser, gravage, livrage, mesurette*). Nous avons consulté le Wiktionnaire : tous y figurent. Nous illustrons en section 4.4.3 comment, dans le cadre d'une description linguistique, le Wiktionnaire permet de compléter les dictionnaires traditionnels. Ainsi, la large couverture et la « réactivité » des ressources collaboratives est donc un atout pour la description linguistique à mettre en balance avec une moindre rigueur éditoriale.

3.2 Description[8]

Wiktionary est un dictionnaire multilingue libre et accessible en ligne. Reposant sur le principe des *wikis*, tout internaute peut le modifier et l'alimenter, les modifications étant publiées immédiatement. Lancé en 2003, ce projet lexicographique fait état, dix ans plus tard, de plus de deux millions d'articles pour son édition française, le *Wiktionnaire*. La taille de cette nomenclature doit être relativisée par le fait qu'elle comptabilise les formes fléchies (notamment les formes conjuguées), des pages de discussion, et, plus curieusement, *« des pages décrivant en français des mots d'autres langues »*. Ces derniers mis à part, le Wiktionnaire contient 1,4 millions d'entrées pour 186 000 lemmes. Il comporte certes quelques excentricités néologiques telles que *nanipabulophiliste*, *énabler* ou *refansubber*. La taille du Wiktionnaire s'explique également par les critères d'acceptation des articles[9] :

> *« La rareté ou la notoriété d'un mot n'est pas un critère, contrairement à Wikipédia, pourvu que l'on soit sûr qu'il soit bien attesté. Les mots récents, les mots familiers, les mots argotiques, sont considérés comme faisant partie de la langue. Il est d'autant plus utile qu'ils sont souvent absents des autres dictionnaires. Les mots et les définitions désuets peuvent figurer sur le Wiktionnaire, du moment qu'ils sont attestés. Par exemple,* loix*, l'ancien pluriel de* loi *n'est jamais utilisé en français*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

> *contemporain, mais existe toujours dans des documents écrits. Les variantes orthographiques d'un même mot peuvent être toutes présentes, sauf s'il s'agit clairement de fautes d'orthographe. »*

Wiktionnaire a bénéficié de l'import d'articles de dictionnaires passés dans le domaine public, tels que la *8e édition du Dictionnaire de l'Académie Française* et, dans une moindre mesure, du *Littré*. Il connaît aujourd'hui une croissance constante grâce à l'édition manuelle des contributeurs. Son infrastructure sous-jacente lui permet d'avoir une macrostructure riche, qui comporte un grand nombre de liens : les pages des formes fléchies renvoient vers les formes citationnelles correspondantes ; les formes verbales renvoient vers des tables de conjugaison ; les liens interlangues, présents dans les sections de traduction, renvoient aux différentes éditions de langue de Wiktionary ; les liens hypertextes présents dans les définitions et les sections relatives aux relations sémantiques renvoient aux entrées correspondantes, etc. Chaque article dispose d'une page web ayant une adresse propre. Un article correspond à une forme graphique. Il comprend une section *Étymologie* suivie de sections *« Type de mot »* (*i.e.* : partie du discours). Chacune de ces sections est numérotée en cas de dégroupement homonymique. La microstructure des sections *« Type de mot »* est composée de définitions et d'exemples pour chacun des sens recensés, et, éventuellement, de relations sémantiques (synonymes, antonymes, hyperonymes et hyponymes), de traductions, de dérivés morphologiques et de « mots apparentés[10] ». On trouve enfin des transcriptions phonémiques dans la plupart des pages. Certaines entrées disposent de plusieurs transcriptions qui rendent compte de leurs variantes diatopiques.

3.3 Critiques

Si nous acceptons le Wiktionnaire pour ce qu'il est, c'est-à-dire à la fois un dictionnaire écrit en grande partie par des amateurs (donc potentiellement entaché d'erreurs et d'inexactitudes), mais également la seule ressource répondant aux besoins énoncés plus haut (cf. section 2), nous pouvons toutefois formuler deux critiques. Tout d'abord, les wikis, souvent présentés comme une démocratie de la connaissance où chaque décision est discutée par une grande variété de contributeurs, fonctionne essentiellement sur la base de *« contributeurs actifs »* (contributeurs réalisant au moins cinq éditions par mois) qui représentent une communauté d'une centaine de personnes seulement. Or ces contributeurs sont souvent également investis dans la gestion du Wiktionnaire (chaque wiki est un écosystème dans lequel évoluent administrateurs, bureaucrates, patrouilleurs, robots et leurs « dresseurs », etc.). La consultation des pages de discussion liées à chaque article, ainsi que les pages de la *Wikidémie* (sorte de forum de discussion), révèle que les *« wiktionnaristes »* sont des contributeurs de bonne volonté que les heures de labeur bénévole rendent peu enclins à se faire déposséder du fruit de leur travail. Dans ce contexte, le nouveau-venu, fût-il expert du domaine, aura moins de crédit que l'amateur hyperactif. Ainsi, certaines erreurs pourtant relevées perdureront faute pour le relecteur vigilant d'avoir su convaincre la majorité (ou d'en avoir pris le temps).

Une autre critique que l'on peut formuler, plus sévère, tient au choix de l'infrastructure logicielle qui supporte le Wiktionnaire, et plus généralement toutes les éditions de Wiktionary : les articles d'un wiki sont stockés sous forme de texte libre, plus ou moins formaté typographiquement. Si ce format peut convenir pour une encyclopédie, il nous paraît aberrant d'avoir initié en 2003 un projet de dictionnaire faisant l'impasse sur une structuration en base de données. Nombre d'erreurs et d'incohérences en découlent et s'ajoutent aux erreurs de jugement des contributeurs. En effet, si des conventions existent, aucun système de saisie ne contraint le rédacteur d'un article, ni ne vérifie la cohérence ou la redondance d'informations. Ce gestionnaire de contenu minimaliste n'autorise aucune systématisation du formatage des pages. Nous montrons dans la section 4.1.2 les implications de ce format non structuré sur un processus d'extraction d'informations. Notons que cette insatisfaction est partagée par d'autres et constitue le point de départ d'OmegaWiki[11], un projet dont le but est *« de créer un dictionnaire de tous les mots de toutes les langues, incluant des informations lexicales, terminologiques et ontologiques »*. On peut lire également que *« [les] données sont stockées dans une base de données relationnelles ce qui permet de les réutiliser facilement »*. Si ces intentions sont louables, l'attractivité d'un tel site est proportionnelle à la masse des informations qu'il contient. Il est à craindre que les internautes aient plus tendance à graviter autour de Wiktionary qu'à contribuer à un projet né après lui, dont il peut leur sembler constituer un doublon.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

4 Des lexiques électroniques fondés sur le Wiktionnaire

GLÀFF, le lexique que nous avons construit à partir du Wiktionnaire, comprend 1,4 millions d'entrées et fournit pour chacune d'elle un lemme, une description morphosyntaxique et, dans 90% des cas, une ou plusieurs transcriptions phonémiques[12]. Placé sous licence libre, ce lexique est téléchargeable depuis le site REDAC[13]. Après avoir décrit en section 4.1.2 comment nous avons construit ce lexique, nous en donnons en sections 4.2 et 4.3 une caractérisation de sa couverture (nombre de lemmes et de formes, couverture relativement à différents corpus) et de ses transcriptions phonémiques. Nous comparons GLÀFF à quatre lexiques utilisés dans de nombreuses recherches : Lexique, BDLex, Morphalou et Lefff. Nous apportons ainsi des éléments de réponse aux questions suivantes : Que contient ce lexique ? Quel est l'apport de ce lexique relativement aux ressources similaires existantes ? Ce lexique est-il une ressource susceptible de remplacer les lexiques morphosyntaxiques et phonologiques les plus couramment utilisés ?

GLÀFF, conçu principalement pour l'exploration linguistique et le TAL, contient des informations pertinentes pour d'autres champs. Nous présentons en section 4.4.1 PsychoGLÀFF, un lexique orienté vers la psycholinguistique construit à partir de GLÀFF. Tout comme GLÀFF, PsychoGLÀFF est distribué sous licence libre. La taille de ces deux ressources les rendant difficilement manipulables autrement que par programme, nous avons conçu GLÀFFOLI, une interface d'interrogation en ligne que nous présentons en section 4.4.2. Enfin, nous donnons en section 4.4.3 quelques exemples typiques d'exploitation que l'on peut réaliser simplement à partir des lexiques issus du Wiktionnaire.

4.1 GLÀFF, *« un Gros Lexique À tout Faire du Français »*

4.1.1 Travaux antérieurs

Le potentiel du Wiktionnaire en tant que lexique électronique a été étudié pour la première fois par Navarro et al. (2009) pour le français et l'anglais. Des travaux ont suivi cette initiative pour d'autres langues. Anton Pérez et al. (2011) décrivent par exemple l'intégration de l'édition portugaise de Wiktionary dans l'ontologie Onto.PT (Gonçalo Oliveira et Gomes, 2010). Citons également Dbnary (Sérasset, 2012), une ressource et un extracteur *open source* visant à extraire de Wiktionary un réseau multilingue. L'auteur précise que ce travail ne vise pas l'exhaustivité mais la conception d'un modèle simple permettant de représenter autant de données qu'il est possible d'extraire correctement, laissant de côté certaines structures pour faciliter cette extraction. Le réseau extrait possède 260 467 entrées pour le français. Le laboratoire UKP distribue deux ressources issues de Wiktionary : OntoWiktionary (Meyer et Gurevych, 2012a), une ontologie construite semi-automatiquement et UBY (Gurevych et al., 2012), un alignement de sept ressources lexicales, disponible pour l'allemand et l'anglais. Si la version allemande de Wiktionary semble être celle qui bénéficie de l'encodage le plus rigoureux et le plus systématique[14], la version française, moins aisément exploitable, se distingue par une plus grande nomenclature, ainsi que la présence quasi-systématique d'informations flexionnelles et phonémiques. Nous avons mis à disposition WiktionaryX[15], une version structurée au format XML de ce dictionnaire pour le français et l'anglais (Sajous et al., 2010).

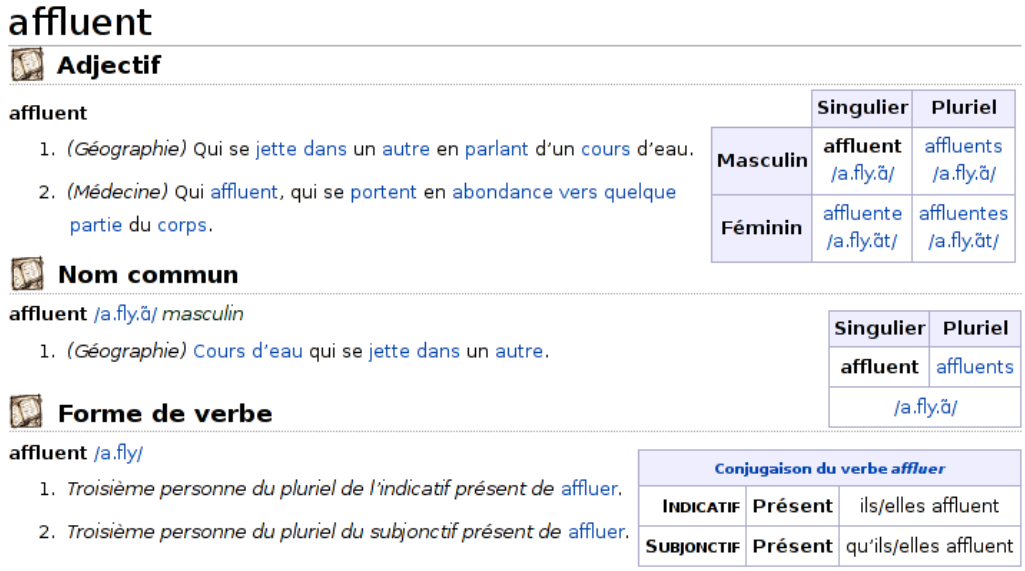
4.1.2 Extraction d'information

Pour chaque édition de langue de Wiktionary, l'ensemble des articles est régulièrement mis à disposition sous forme de fichiers appelés *XML dumps*[16]. Il ne faut pas interpréter la mention « XML » comme l'encodage de la mircrostructure des articles par des balises XML. Les balises ne servent qu'à délimiter les articles et leur titre. Le reste du contenu est encodé dans un format appelé *wikicode*, inhérent au système de gestion de contenu *MediaWiki*. La syntaxe de ce format n'a jamais été définie formellement et, de plus, évolue dans le temps, avec coexistence de plusieurs conventions d'encodage pour un même type d'information. Notons également que ni les conventions d'organisation des articles, ni leur encodage en *wikicode*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

n'est stable d'une édition de langue à l'autre.

Ce format lâche rend ardue et constamment inachevée l'écriture d'un parseur pour extraire de manière automatique et exhaustive les informations du Wiktionnaire (Navarro et al., 2009; Sajous et al., 2010, 2011). Entre la mise à disposition de deux *dumps*, le wikicode évolue sans que le changement ne soit nécessairement documenté et seule l'observation (semi-)manuelle du format d'encodage permet d'adapter le parseur en conséquence. Le travail présenté ici porte principalement sur l'extraction des informations flexionnelles et des transcriptions phonémiques.



affluent

Adjectif

affluent

1. *(Géographie)* Qui se jette dans un autre en parlant d'un cours d'eau.
2. *(Médecine)* Qui affluent, qui se portent en abondance vers quelque partie du corps.

	Singulier	Pluriel
Masculin	**affluent** /a.fly.ɑ̃/	affluents /a.fly.ɑ̃/
Féminin	affluente /a.fly.ɑ̃t/	affluentes /a.fly.ɑ̃t/

Nom commun

affluent /a.fly.ɑ̃/ *masculin*

1. *(Géographie)* Cours d'eau qui se jette dans un autre.

Singulier	Pluriel
affluent	affluents
/a.fly.ɑ̃/	

Forme de verbe

affluent /a.fly/

1. *Troisième personne du pluriel de l'indicatif présent de* affluer.
2. *Troisième personne du pluriel du subjonctif présent de* affluer.

Conjugaison du verbe *affluer*		
Indicatif	Présent	ils/elles affluent
Subjonctif	Présent	qu'ils/elles affluent

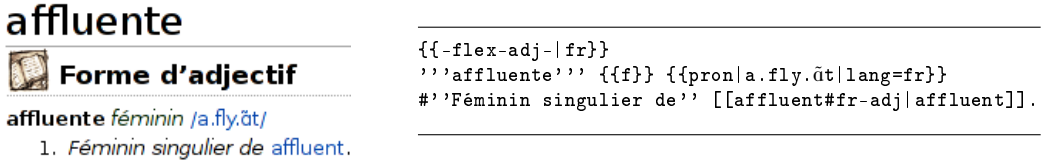
(a) Aperçu de l'article *« affluent »*

```
{{-adj-|fr}}
{{fr-accord-cons|a.fly.ɑ̃|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]]
d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ɑ̃}}

{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui|}}
'''affluent''' {{pron|a.fly|fr}}
# ''3ème pers. du pluriel de l'indicatif présent de'' [[affluer]].
# ''3ème pers. du pluriel du subjonctif présent de'' [[affluer]].

{{-pron-}}
{| class="wikitable"
| Adjectif et nom commun
* {{pron-rég|France|ɛ̃ .n‿a.fly.ɑ̃|titre=un affluent}}
|-
| Forme du verber affluer
* {{pron-rég|France (Île-de-France)|a.fly}}
```

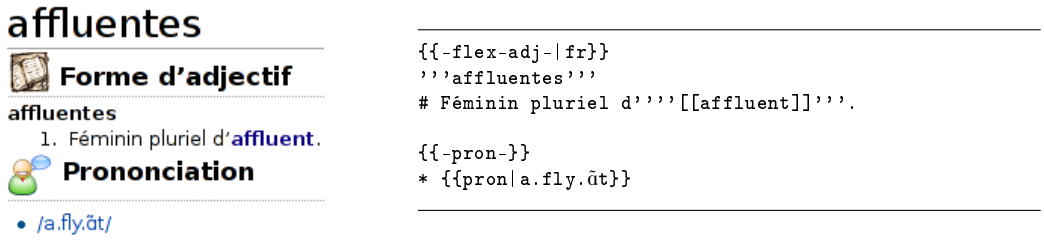
(b) Wikicode de l'article *« affluent »*

Figure 1 – Article *« affluent »* dans le Wiktionnaire et wikicode correspondant.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216



(a) Article *« affluente »* et wikicode correspondant



(b) Article *« affluentes »* et wikicode correspondant

Figure 2 – Exemples de formes fléchies d'*« affluent »* dans le Wiktionnaire.

Les figures 1 et 2 montrent des extraits de l'article *« affluent »*, tel qu'on peut le consulter dans le Wiktionnaire, des articles de deux de ses formes fléchies, ainsi que le wikicode correspondant à chaque extrait. Le tableau qui recense les formes fléchies de l'adjectif et les transcriptions phonémiques correspondantes (en haut à droite de la figure 1a) n'est pas explicitement présent dans le wikicode, mais généré par le patron `{{fr-accord-cons|a.fly.ã|t}}` (figure 1b). Il existe plusieurs dizaines de patrons similaires dans le wikicode. L'extraction des formes fléchies et des prononciations correspondantes se fait soit par recensement et « émulation » de ces patrons paramétrables (ici, génération des formes fléchies à partir d'un schéma spécifié), soit par analyse des articles des formes fléchies lorsqu'ils existent (cf. fig. 2a et 2b). Là encore, aucun formatage n'est systématique : le patron `{{f}}` (fig. 2a) indique que la forme est de genre féminin ; le nombre doit être extrait du texte de la ligne suivante *« Féminin singulier »*. Si la prononciation de *affluente* est donnée dans la « ligne de forme », celle de *affluentes* est donnée dans une section *Prononciation* dédiée. Des erreurs induites par l'hétérogénéité du wikicode peuvent de ce fait s'ajouter aux erreurs contenues dans les articles du Wiktionnaire et avoir un impact sur la ressource finale. Notre parseur extrait du Wiktionnaire les formes graphiques et leurs lemmes, convertit leurs catégories morphosyntaxiques au format GRACE (Rajman et al., 1997) et extrait leurs transcriptions phonémiques, données en API. En cas de transcriptions multiples, comme /ply/ et /plys/ pour l'adverbe *plus*, toutes sont conservées. Les informations flexionnelles présentes dans le Wiktionnaire sont parfois incomplètes. Il est en effet courant que seul le genre ou le nombre soit indiqué pour les noms et les adjectifs. De même, il arrive que l'information pour les formes verbales fléchies omette le temps ou le mode. Nous appliquons des règles pour tenter de compléter ces informations : le genre et le nombre d'un participe passé peuvent être inférés par sa terminaison ; une forme fléchie nominale ou adjectivale masculine, dont on a déjà rencontré le lemme masculin singulier, est plurielle ; etc. Les 9,5% d'entrées dont l'information flexionnelle reste partielle à l'issue de la complétion sont écartées de la ressource. Un extrait de GLÀFF est donné en figure 3. On y voit qu'une entrée est composée d'une forme graphique, d'une catégorie morphosyntaxique, d'un lemme et de transcriptions phonémiques en API et en SAMPA. Dans la version actuelle de GLÀFF, seuls sont inclus les mots grammaticaux, noms communs, verbes, adjectifs et adverbes. Les locutions y seront intégrées ultérieurement. En complément du *dump* dont elles sont absentes, nous avons « aspiré » du site du Wiktionnaire, puis analysé, les tables de conjugaison de 18 076 verbes. Ces tables générées à partir d'un simple modèle (e.g. `{{fr-conj-1|march|pron=maʁ|pc=ʃ}}` pour le verbe *marcher*[17]) permettent d'obtenir les 48 formes fléchies simples d'un verbe.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216


```
affluent|Afpms|affluent|a.fly.ã|a.fly.A~
affluente|Afpfs|affluent|a.fly.ãt|a.fly.A~t
affluent|Ncms|affluent|a.fly.ã|a.fly.A~
affluents|Afpmp|affluent|a.fly.ã|a.fly.A~
affluents|Ncmp|affluent|a.fly.ã|a.fly.A~
affluent|Vmip3p-|affluer|a.fly|a.fly
```

Figure 3 – Extrait de GLÀFF

4.2 Couverture

GLÀFF se distingue des lexiques disponibles essentiellement utilisés en TAL et en psycholinguistique par sa taille exceptionnelle. Le tableau 1 présente le nombre de lemmes et de formes fléchies, simples (séquences de lettres exclusivement) et non simples (*i.e.* comprenant espace, tiret et/ou chiffre). GLÀFF contient 3 à 4 fois plus de lexèmes et 3 à 9 fois plus de formes que les autres lexiques. Cette taille est un atout important pour certaines utilisations, notamment les recherches en morphologie flexionnelle ou dérivationnelle. Elle est également intéressante pour le développement d'outils de TAL comme des étiqueteurs morphosyntaxiques ou des analyseurs syntaxiques. GLÀFF comporte aussi un nombre élevé de formes composées. Ces dernières servent essentiellement à la segmentation des textes dont dépend directement la qualité des annotations catégorielles et syntaxiques ultérieures.

	Formes fléchies catégorisées			**Lemmes catégorisés**		
	Simples	Non simples	Total	Simples	Non simples	Total
Lexique	147 912	4 696	152 608	46 649	3 770	50 419
BDLex	431 992	4 360	436 352	47 314	1 792	49 106
Lefff	466 668	3 829	470 497	54 214	2 303	56 517
Morphalou	524 179	49	524 228	65 170	7	65 177
GLÀFF	1 401 578	24 270	1 425 848	172 616	13 466	186 082

Tableau 1 – Taille des lexiques (restreints aux catégories : nom commun, verbe, adjectif, adverbe).

Les comparaisons ci-après concernent uniquement les catégories nom commun, verbe, adjectif et adverbe. Elles ont été réalisées sur les formes graphiques ou lemmes « simples » afin de nous affranchir des différents choix de graphie des unités polylexicales dans les lexiques et de segmentation dans les corpus[18]. Nous étudions tout d'abord l'intersection de GLÀFF avec les quatre autres lexiques. Le tableau 2 présente pour chacun des cinq lexiques testés la proportion d'entrées (*i.e.* de triplets *<forme; lemme; description morphosyntaxique>*) que l'on retrouve dans les autres. On observe que la taille des intersections est directement liée à celle des lexiques : plus un lexique est gros, plus son intersection avec les autres l'est. On observe ensuite une répartition des cinq lexiques en trois groupes : Lexique a une couverture moindre, avec 9% de celle de GLÀFF et 22 à 26% de celle des lexiques BDLex, Lefff et Morphalou. Ces trois derniers couvrent 76% à 80% de Lexique et 30% de GLÀFF en moyenne, tout en ayant une couverture commune de 70% à 86%. GLÀFF est nettement au-dessus avec une couverture de 85% à 93%. Sa couverture est supérieure de 5% à 65% à celle des autres lexiques. GLÀFF a donc une taille nettement supérieure à celle des autres lexiques, ce qui constitue un atout potentiel. Afin de s'assurer que cet avantage est effectif (*i.e.* que le plus grand nombre de lexèmes et de formes peut réellement s'avérer utile), nous avons comparé les cinq lexiques au vocabulaire de quatre corpus de nature différente (genre, taille, époque, etc.). Le premier, composé de 515 romans du XX^{e} siècle issus de la base Frantext[19], contient 30 millions de mots. LM10, corpus journalistique qui rassemble les archives de 1991 à 2000 du quotidien *Le Monde*, contient 200 millions de mots. Le troisième corpus, composé des 664 982 articles de la Wikipédia française[20], contient 260 millions de mots. Enfin, FrWaC (Baroni et al., 2009) est un corpus de pages Web en français (plus précisément, du domaine `.fr`) contenant 1,6 milliard de mots. Ces quatre corpus ont été étiquetés par la version standard

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

	Lexique	**BDLex**	**Lefff**	**Morphalou**	**GLÀFF**
Lexique		26,03	25,20	22,46	8,95
BDLex	76,02		79,87	70,40	28,75
Lefff	79,50	86,28		72,32	30,04
Morphalou	79,58	85,43	81,24		32,03
GLÀFF	**84,83**	**93,26**	**90,23**	**85,66**	

Tableau 2 – Couverture inter-lexiques (en % de formes fléchies catégorisées).

de TreeTagger[21], qui nous sert ici seulement à segmenter les textes et filtrer le vocabulaire sur la base des catégories syntaxiques. Les mots inconnus de TreeTagger (dont la catégorie est pertinente) sont conservés. Le tableau 3 présente la couverture des cinq lexiques par rapport à ces quatre corpus, en distinguant au sein de leur vocabulaire les formes de fréquence supérieure ou égale à 1 (*i.e.* tout le vocabulaire), 2, 5, 10, 100 et 1000. Le classement des corpus par couverture décroissante est le même pour les cinq lexiques. Bien que la taille des corpus influe sur cet ordre (plus un corpus est étendu, plus le nombre potentiel de formes différentes est grand), leur nature est également déterminante : FrWaC, par exemple, est une collection de pages web et (donc) contient nombre de formes « bruitées » (mots étrangers, espaces manquants ou excédentaires, orthographe aléatoire, absence de diacritiques, etc.). On retrouve la répartition des lexiques en trois groupes : BDLex, Lefff et Morphalou présentent une couverture assez proche. Hormis pour Frantext, Lexique affiche une couverture moindre jusqu'au seuil 100, où il rejoint Morphalou. GLÀFF a une couverture supérieure pour les trois plus gros corpus, sauf pour LM10 au seuil 1000 où il est dépassé par Lefff de 0,2%. La meilleure couverture de Lexique pour Frantext, alors qu'elle est inférieure de 10 à 15% à celle de GLÀFF pour les trois autres corpus, s'explique probablement par le fait que son vocabulaire a été constitué à partir d'œuvres de cette même base. Pour les autres corpus et jusqu'au seuil 100, la taille de GLÀFF lui permet d'avoir une couverture du vocabulaire bien supérieure à celle des autres lexiques (au seuil 1, de 14% à 53% de plus pour LM10 et de 30% à 125% pour FrWaC ; au seuil 10, de 4% à 16% pour LM10 et de 15% à 47% pour FrWaC). De plus, le fait qu'il n'inclue que partiellement les autres lexiques n'est pas surprenant au vu des intersections de ces derniers.

La figure 4 compare la couverture des cinq lexiques sous un autre éclairage : elle représente pour chacun le nombre de formes dont la fréquence en corpus appartient à un intervalle donné. On y voit clairement que les différences sont plus marquées pour le corpus FrWaC qu'elles ne le sont pour Frantext, probablement du fait des différences liées à la nature des corpus, comme expliqué plus haut. La répartition des lexiques en trois groupes apparaît clairement dans le diagramme de droite (FrWaC). On voit également dans ce dernier que même pour les mots très fréquents et donc très bien attestés, qui ont par exemple une fréquence comprise entre 101 et 1000, la couverture de GLÀFF reste meilleure. Le tableau 3 et la figure 4 montrent que la supériorité de GLÀFF est plus marquée lorsque l'on travaille sur des corpus hétérogènes et pour des mots de faibles et moyennes fréquences.

Pour conclure notre caractérisation de la couverture de GLÀFF, nous nous sommes intéressés à la partie du vocabulaire spécifique à ce dernier, *i.e.* aux formes appartenant à GLÀFF et absentes des quatre autres lexiques. Ce sous-ensemble de 665 290 formes représente 47% de la ressource. Nous avons également considéré le vocabulaire spécifique de chaque autre lexique. Le tableau 4 montre pour chaque sous-vocabulaire le nombre de formes attestées en corpus. Conformément à l'intuition, le nombre de formes attestées est d'autant plus grand que les corpus sont gros. La taille des corpus n'explique cependant pas tout : si une large part du vocabulaire spécifique à GLÀFF n'est attestée dans aucun corpus (il s'agit majoritairement de verbes pour lesquels toutes les flexions possibles sont générées), sa taille permet une meilleure couverture de corpus hétérogènes tels que FrWaC incluant un français potentiellement moins normé et plus récent. Même pour un corpus journalistique dont l'année la plus récente est 2000, la jeunesse, mais également la mise à jour constante du Wiktionnaire permettent à GLÀFF de couvrir des mots tout à fait usuels comme : *transversalité*, *attractivité*, *brevetabilité*, *diabolisation*, *employabilité*, *anticorruption*, *homophobie*, *institutionnellement*, *hébergeur*, *fatwa*, *indétrônable*, *recapitalisation*, *autofiction*, *déremboursement*,

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

relégable, *sanctuarisation*, *préretraité*, etc., qui restent, et resteront, absents des autres lexiques, faute de mise à jour.

Seuil : fréquence ≥		**1**	**2**	**5**	**10**	**100**	**1000**
Frantext	Nb formes	1 45 437	95 189	61 813	43 919	10 767	1 376
	Lexique	66,76	84,35	94,00	**96,91**	**99,15**	**99,27**
	BDLex	70,86	84,69	92,47	95,74	99,12	99,20
	Lefff	71,89	85,63	93,21	96,21	99,08	98,90
	Morphalou	73,93	86,66	93,29	96,00	98,48	97,09
	GLÀFF	**76,92**	**88,57**	**94,54**	96,72	98,77	98,76
LM10	Nb formes	300 606	172 036	106 470	77 936	29 388	7 838
	Lexique	29,59	47,28	65,23	76,31	93,81	98,58
	BDLex	37,77	55,79	71,76	80,93	95,53	98,69
	Lefff	39,64	58,22	74,33	83,20	95,99	**98,90**
	Morphalou	39,06	56,82	71,92	80,32	93,27	97,48
	GLÀFF	**45,24**	**63,83**	**78,63**	**86,23**	**96,46**	98,68
Wikipédia	Nb formes	953 920	435 031	216 210	136 531	35 621	7 956
	Lexique	9,13	18,27	31,52	43,03	78,58	95,72
	BDLex	12,29	22,89	36,80	48,04	79,39	95,33
	Lefff	12,88	23,94	38,26	49,65	80,57	95,71
	Morphalou	13,05	23,96	37,87	48,87	78,74	94,16
	GLÀFF	**16,42**	**29,00**	**44,13**	**55,45**	**83,21**	**96,10**
FrWaC	Nb formes	1 624 620	846 019	410 382	255 718	74 745	22 100
	Lexique	5,83	10,85	20,84	30,81	66,00	89,47
	BDLex	9,36	15,85	27,28	37,48	69,61	90,03
	Lefff	9,85	16,67	28,57	39,16	71,61	91,16
	Morphalou	10,09	16,89	28,53	38,68	69,36	88,51
	GLÀFF	**13,13**	**21,13**	**34,29**	**45,35**	**76,39**	**92,76**

Tableau 3 – Couverture lexiques/corpus (en % de formes fléchies non catégorisées).

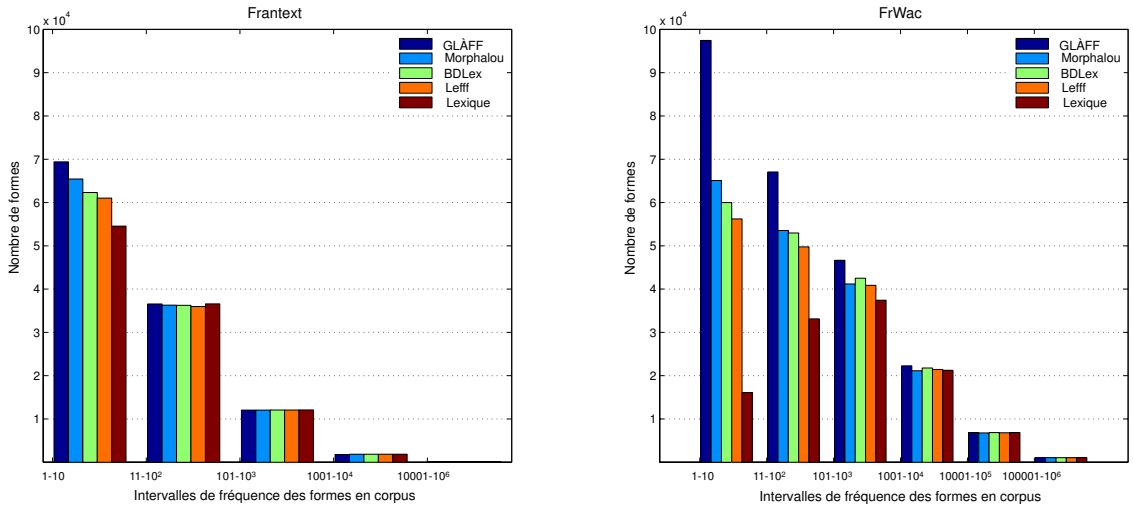


Figure 4 – Répartition des formes des lexiques relativement à leur fréquence en corpus.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

	Taille du vocabulaire spécifique	Nombre de formes attestées			
		Frantext	LM10	Wikipédia	FrWaC
Lexique	1 509	866	863	1 073	1 320
BDLex	3 981	86	521	1 004	1 496
Lefff	11 050	232	1 479	2 214	3 288
Morphalou	26 881	1 171	1 912	3 995	6 425
GLÀFF	665 290	2 811	13 525	29 230	47 549

Tableau 4 – Attestation en corpus du vocabulaire spécifique de chaque lexique

4.3 Transcriptions phonémiques

GLÀFF contient, pour 90% de ses entrées, une transcription phonémique. Ces transcriptions contiennent plusieurs variantes dans 8% des cas. Afin d'évaluer leur qualité, nous les avons comparées à celles de BDLex et de Lexique, que nous avons converties en API. Nous avons comparé d'une part les transcriptions sans tenir compte de la syllabation, puis nous avons comparé la syllabation pour les transcriptions dont les suites de phonèmes sont strictement identiques. Nous avons également relevé, pour les transcriptions qui ne diffèrent que par un phonème, les oppositions en cause. Le tableau 5 montre pour chaque couple de lexiques les dix oppositions les plus fréquentes et le tableau 6 donne des exemples illustrant ces oppositions. Ce tableau est complété, dans la dernière colonne, par les transcriptions du *Dictionnaire de la Prononciation Française dans son Usage Réel* (Martinet et Walter, 1973), noté DPF ci-après. Les auteurs de ce dictionnaire papier élaboré entre 1968 et 1973, partant du principe que *« l'unité de la prononciation française est une vue de l'esprit et ne correspond à rien de réel »* ont mené, avec leurs collaborateurs, un travail de recensement auprès de 17 informateurs pour collecter les différentes variantes de prononciation d'un même mot (20% des prononciations divergent). Les différences de transcription entre GLÀFF et chacun des deux autres lexiques sont comparables aux différences que l'on trouve entre BDlex et Lexique. Elles sont principalement dues à l'opposition entre les voyelles moyennes, comme les antérieures : [e] (mi-fermée) vs. [ɛ] (mi-ouverte), et les postérieures : [o] (mi-fermée) vs. [ɔ] (mi-ouverte). Entre BDLex et Lexique, elles sont responsables de 91% des divergences. Ces différences étaient attendues : l'opposition entre voyelles mi-fermeés et mi-ouvertes en français est soumise à des restrictions distributionnelles définies par la *loi de position* selon laquelle les voyelles mi-ouvertes apparaissent de préférence en syllabe fermée, alors que les voyelles mi-fermées apparaissent de préférence en syllabe ouverte. Les autres oppositions relevées dans le tableau 5, comme l'opposition [s]/[z], venant principalement du suffixe *-isme*, sont décrites dans le DPF. Le codage problématique du schwa y est également longuement commenté.

Le tableau 7 montre la proportion de transcriptions strictement identiques (hors syllabation), et « comparables » après annulation des différences entre voyelles moyennes. Cette notion de « comparabilité », ainsi arbitrairement définie, montre que la majorité des différences (97 à 98%) sont dues à ces seules oppositions et ne viennent pas de codages aberrants. GLÀFF et Lexique proposent des prononciations strictement identiques pour 79,5% des entrées. Cet accord strict est de 61,7% entre GLÀFF est BDLex. On peut donc estimer que les transcriptions phonémiques de GLÀFF sont de bonne qualité (l'accord entre BDLex et Lexique n'est que de 58,3%). Notons également que les emprunts sont souvent générateurs de divergences (e.g. *shaker* : /ʃɛi.kəʁ/, /ʃɛj.kœʁ/, /ʃɛ.kœʁ/; *chili* : /ʃi.li/, /tʃi.li/; *ginseng* : /ʒin.sɑ̃g/, /ʒin.sɑ̃ŋ/, /ʒin.sɛŋ/). Par ailleurs, ni Lexique ni BDLex ne saurait constituer un étalon absolu. Si l'opposition [o]/[ɔ] peut s'expliquer, certaines entrées transcrites avec un [o] (o fermé) dans BDLex sont surprenantes : /po.m/ pour *pomme*, /poʁt/ pour *porte*, /oʁ/ pour *or* et *hors*, etc. Concernant Lexique, que penser de *châtié*, transcrit /ʃa.sje/, ou de *cambriolé/cambriolés* transcrits respectivement /kɑ̃.bʁi.jo.le/ et /kɑ̃.bʁi.o.le/ ? On s'étonne également de lire dans sa documentation[22] que le caractère `9` code le « e-ouvert [comme dans] *œuf, peur* » et de trouver dans le lexique *peur* transcrit `/p2ʁ/`, `2` étant selon la documentation le code pour le « e-fermé [comme dans] *deux* ». Une autre curiosité concerne le « schwa non élidable [comme dans] *parvenu* », codée selon la documentation par le symbole `3`. Or ce symbole est totalement absent de Lexique. *Parvenu*, utilisé comme

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

Op.	Phonèmes	%	% cumulé
r	ɛ/e	48,18	48,18
r	ɔ/o	32,17	80,36
r	o/ɔ	11,02	91,37
r	y/ɥ	1,83	93,21
r	ə/ø	1,44	94,64
r	ə/œ	1,39	96,03
r	u/w	0,84	96,87
r	b/p	0,73	97,61
r	s/z	0,51	98,12
d	j	0,25	98,37

(a) BDLex/Lexique

Op.	Phonèmes	%	% cumulé
r	ɔ/o	60,03	60,03
i	ə	14,18	74,21
r	e/ɛ	6,90	81,11
r	ɛ/e	4,98	86,09
r	ɑ/a	4,92	91,01
r	s/z	1,25	92,26
r	ə/ø	0,91	93,17
r	œ/ø	0,47	93,64
i	i	0,42	94,06
r	o/ɔ	0,38	94,44

(b) GLÀFF/Lexique

Op.	Phonèmes	%	% cumulé
r	e/ɛ	66,46	66,46
r	ɔ/o	10,58	77,05
i	ə	5,90	82,96
r	o/ɔ	4,36	87,32
r	ɑ/a	3,84	91,17
r	ɥ/y	1,61	92,78
r	œ/ə	1,09	93,88
r	ø/ə	0,86	94,74
i	i	0,84	95,58
r	w/u	0,79	96,38

(c) GLÀFF/BDLex

Tableau 5 – Les 10 différences de transcription les plus fréquentes.
Opérations (Op.) : r = substitution ; i = insertion ; d = suppression.

		Transcriptions			
Opération	**Forme**	**BDLex**	**Lexique**	**GLÀFF**	**DPF**
r : ɛ/e	été	/ɛ.te/	/e.te/	/e.te/	/ete/
r : s/z	stalinisme	/sta.li.nis,m/	/sta.li.nizm/	/sta.li.nism/	/stalinism/, /stalinizm/
r : b/p	obturer	/ɔb.ty.ʁe/	/ɔp.ty.ʁe/	/ɔp.ty.ʁe/	/ɔptyre/, /ɔbtyre/
r : o/ɔ	pomme	/po,m/	/pɔm/	/pɔm/	/pɔm/
r : ə/ø/œ	heureux	/ə.ʁø/	/ø.ʁø/	/œ.ʁø/	/ørø, œrø/
r : y/ɥ	gradué	/gʁa.dy.e/	/gʁa.dɥe/	/gʁa.dɥe/	/gradɥe/, /grɑdɥe/, /gradye/
r : u/w	jouer inouï	/ʒu.e/ /i.nu.i/	/ʒwe/ /i.nwi/	/ʒwe/ /i.nwi/	/ʒwe/, /ʒue/ /inwi/, /inui/
r : a/ɑ	pâte	/pa,t/	/pat/	/pɑt/	/pat/ , /pɑt/
i,d : i,j	riiez	/ʁi.i.je/	/ʁi.je/	/ʁij.je/	-
i,d : ə	contenu	/kɔ̃,tə.ny/	/kɔ̃.tə.ny/	/kɔ̃t.ny/	/kɔ̃t(ə)ny/

Tableau 6 – Exemples de différence de transcription entre lexiques.

			Transcriptions phonémiques		Syllabation
Lexiques		**Intersection**	**Identiques**	**Comparables**	**Identiques**
BDLex	Lexique	112 439	58,31	96,88	98,92
GLÀFF	Lexique	123 630	79,50	97,81	98,48
GLÀFF	BDLex	396 114	61,72	96,88	98,30

Tableau 7 – Accord inter-lexiques : transcriptions phonémiques et syllabation
(Transcriptions comparables : non prise en compte des oppositions [o]/[ɔ], [e]/[ɛ] et [œ]/[ə]/[ø].
Syllabation : comparaison sur les transcriptions phonémiques identiques)

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

exemple, est transcrit /paRv˚ny/, où ˚ code le schwa élidable.
Une étude quantitative et qualitative des transcriptions phonémiques de Wiktionary avait été menée par Schlippe et al. (2010) pour les éditions française, anglaise, espagnole et allemande. Cette étude montrait que le Wiktionnaire avait le plus fort pourcentage d'articles comportant une transcription (51%[23] contre 9% pour l'édition anglaise, par exemple). Ces transcriptions ont été comparées à celles GlobalPhone (cf. section 2), générées par un système à base de règles, puis vérifiées manuellement. Là encore, la qualité (estimée de cette manière) des transcriptions a été jugée meilleure pour l'édition française, avec 74% de transcriptions identiques contre respectivement 50%, 28% et 26% pour les versions espagnole, allemande et anglaise.
La comparaison de la syllabation opérée sur les transcriptions identiques (cf. tableau 7) montre que les trois lexiques sont très proches (98%). Notons à ce propos que si la construction « collaborative par les foules » du Wiktionnaire peut dans certains cas être source d'amateurisme, elle peut également être intéressante car elle reflète une perception non canonique de la langue, selon un point de vue qui est *de facto* celui du locuteur (en l'occurrence, le contributeur) et non *de jure* celui du lexicographe. À titre d'exemple, on peut citer le cas de la syllabation du groupe consonantique /s/ + C en position interne de mot. Dans GLÀFF ce groupe apparaît alternativement comme hétérosyllabique, *i.e.* le /s/ et la consonne qui suit appartiennent à deux syllabes différentes (c'est la version canonique en français) comme dans *ministère* /mi.nis.tɛʁ/ et comme tautosyllabique (les deux phonèmes appartiennent à la même syllabe) comme dans *monistique* /mɔ.ni.stik/. Cette alternance, avec d'autres phénomènes non stables dans le Wiktionnaire, peuvent être perçus comme les signaux du comportement parfois non déterministe de la langue, et, partant, comme des objets potentiels d'investigation linguistique et psycholinguistique.

4.4 Autour de GLÀFF

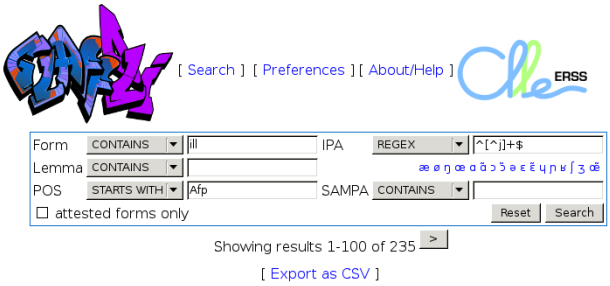
4.4.1 PsychoGLÀFF, un lexique pour la psycholinguistique

Les ressources lexicales jouent un rôle important en psycholinguistique, notamment dans la mise en place d'expériences portant sur l'accès lexical. Dans ce domaine, le lexique CELEX (Baayen et al., 1995) est certainement la ressource (payante) la plus connue et utilisée. Elle fournit pour l'anglais, l'allemand et le néerlandais un ensemble d'informations morphophonologiques validées manuellement. Pour le français, Lexique (New, 2006) est une ressource libre, téléchargeable et interrogeable en ligne. Comme nous l'avons mentionné plus haut, elle présente cependant une couverture limitée.
À partir de GLÀFF, pensé et conçu pour l'exploration linguistique et les applications de TAL, susceptibles de tirer un bénéfice de sa large couverture, nous avons développé PsychoGLÀFF[24], un lexique orienté vers la psycholinguistique. Dans cette discipline, la fréquence et les propriétés distributionnelles d'un mot sont des informations essentielles pour appréhender le répertoire lexical des locuteurs. Dans cette perspective, PsychoGLÀFF est constitué des entrées de GLÀFF attestées dans au moins un des corpus présentés plus haut (Frantext, LM10 et FrWaC), soit environ 330 000 formes pour 121 000 lemmes. Chaque entrée de PsychoGLÀFF fournit les informations suivantes (en plus de celles déjà présentes dans GLÀFF) : la longueur du mot en nombre de caractères et de phonèmes ; sa structure syllabique (annotation CV) et le rapport entre le nombre de consonnes et le nombre de syllabes (ce score est utilisé pour estimer la « complexité syllabiques » des mots) ; des informations relatives à la similarité formelle intra-lexicale, comme le nombre de voisins orthographiques (resp. phonémiques) de chaque mot, *i.e.* le nombre d'entrées du lexique qui ne diffèrent que par un caractère (resp. phonème). PsychoGLÀFF fournit également un score définissant une « régularité phonotactique » (Bailey et Hahn, 2001), calculée à partir de la probabilité des séquences des *n*-grammes contenus dans le mot. Cette mesure est particulièrement utile dans la conception de stimuli pour la mise en place d'expériences portant sur l'accès lexical (Storkel et Hoover, 2011).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

4.4.2 GLÀFFOLI, une interface d'interrogation de GLÀFF

Afin de faciliter l'exploration du volumineux GLÀFF, nous avons conçu GLÀFFOLI (GLÀFF OnLine Interface), une interface d'interrogation en ligne (représentée en figure 5), qui permet d'effectuer des requêtes multicritères portant sur la forme des entrées, leur lemme, leur catégorie syntaxique et leur prononciation. Chaque critère de recherche peut être formulé par une expression régulière ou par un opérateur (*est, contient, commence par, finit par*). L'affichage est paramétrable et permet de visualiser ou non les transcriptions SAMPA et les fréquences en corpus. Lorsqu'une forme est attestée dans le corpus FrWaC, un lien sur sa fréquence pointe vers le concordancier NoSketchEngine (Rychlý, 2007), qui fournit les contextes d'apparition de ces formes dans ce corpus. Enfin, pour chaque requête, l'ensemble des résultats est exportable dans un fichier `CSV` afin de permettre son utilisation dans un tableur tel qu'Excel ou OpenOffice.



					Frantext 20e				LM10				FrWaC			
Form	POS	Lemma	IPA	SAMPA	Form ↓	↑	Lemma ↓	↑	Form ↓	↑	Lemma ↓	↑	Form ↓	↑	Lemma ↓	↑
achilletalonesques	Afpfp	achilletalonesque	a.ʃil.ta.lɔ̃.nɛsk	a.Sil.ta.lO~.nEsk	0	0	0	0	0	0	0	0	0	0	0	0
capillaires	Afpfp	capillaire	ka.pi.lɛʁ	ka.pi.lER	12	0.415	20	0.693	87	0.395	144	0.655	1123	0.895	3019	2.407
capillotractées	Afpfp	capillotracté	ka.pi.lɔ.tʁak.te	ka.pi.lO.tRak.te	0	0	0	0	0	0	0	0	2	0.001	11	0.008
bacillaire	Afpfs	bacillaire	ba.si.lɛʁ	ba.si.lER	1	0.034	2	0.069	2	0.009	2	0.009	36	0.028	44	0.035
ancillaire	Afpfs	ancillaire	ɑ̃.si.lɛʁ	A~.si.lER	10	0.346	25	0.866	10	0.045	25	0.113	66	0.052	128	0.102

Figure 5 – GLÀFFOLI, interface d'interrogation en ligne de GLÀFF
`http://redac.univ-tlse2.fr/glaffoli/`

4.4.3 Autres exemples d'exploitation

Nous avons montré dans les sections précédentes comment le Wiktionnaire, pris dans sa globalité, pouvait servir de fondement pour la construction d'un lexique électronique. Nous avons également évoqué en section 3.1 un recours possible au Wiktionnaire comme complément des autres lexiques pour une étude de la néologie. On peut aussi en extraire des sous-lexiques en utilisant les catégories des articles et les notes de registres, ou de domaines, présents dans les gloses. Le « lexique de la linguistique[25] », par exemple, contient 1270 termes tels que *ambitransitif, allomorphe, apaxique, bilabialisation*, etc.
Un travail récent (Flaux et al., 2014) portant sur le recensement des noms d'humains créateurs (e.g. *sympho-niste, poète, romancier, sculpteur*, etc.) a permis d'enrichir la base NHUMA[26] de 15% (ajout de 90 entrées à une base en contenant initialement 516) par simple recherche de mots amorce dans les définitions du Wiktionnaire, alors que d'autres ressources telles que le TLFi et le Dictionnaire Électronique des Synonymes[27] avaient déjà été exploitées, et qu'un moissonage du Web par l'outil WaliM (Namer, 2003) avait été effectué.

Archer (2009), dont la thèse s'intéresse à l'extraction de collocations et de leurs traductions, propose un état de l'art de la collecte d'informations dans ce domaine. Concernant le Wiktionnaire, il mentionne *« le désordre dans lequel sont présentées les informations [qui] rend difficile leur emploi »*. Espérons que le travail initié ici, qui a vocation à intégrer ultérieurement les unités polylexicales, facilitera ce type d'étude.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

5 Conclusion

Le travail présenté dans cet article est une étude agnostique du Wiktionnaire, entre agoraphobie et « wikipédiholisme[28] ». S'il semble en effet absurde de présenter le Wiktionnaire comme un idéal lexicographique, il nous paraît également peu raisonnable de le mettre au rebut sur l'unique base de préjugés. Motivée par l'absence de ressource électronique satisfaisante comportant des transcriptions phonémiques, notre étude entend porter un éclairage sur les atouts du Wiktionnaire, tout en soulignant ses faiblesses. Notre propos n'est pas de mettre le Wiktionnaire sur le même plan que les dictionnaires de référence, mais de déterminer les champs d'application qui peuvent en tirer profit. Nous avons montré qu'en tant que lexique pris dans sa globalité il était adapté aux applications de TAL, qu'il peut se révéler utile pour l'expérimentation psycholinguistique et qu'il peut apporter une contribution à différentes études linguistiques (néologie, morphologie, etc.). Nous mettons à disposition, sous licence libre, les lexiques GLÀFF et PsychoGLÀFF, ainsi que leur interface d'interrogation, GLÀFFOLI.
Par ailleurs, l'exploitation d'une ressource collaborative comme le Wiktionnaire, reflétant la compétence et la perception de la langue des locuteurs-contributeurs, pourrait apporter une contribution significative à l'analyse des différents aspects de variation à des niveaux de granularité multiples, qui interviennent dans la réalisation lexicale.

Les perspectives d'évolution de GLÀFF sont nombreuses. Nous allons dans un premier temps y intégrer les locutions et unités polylexicales. Nous souhaiterions concevoir un système semi-automatique de détection et de correction d'erreurs (avec validation manuelle). Nous prévoyons également d'évaluer l'apport de ce lexique à des outils de TAL comme l'étiqueteur morphosyntaxique Talismane (Urieli et Tanguy, 2013). Enfin, nous projetons de développer une ressource résultant de l'unification de GLÀFF et d'une version actualisée de WiktionaryX. Aux informations morphosyntaxiques et phonologiques de la première viendront s'ajouter définitions, relations sémantiques lexicales et traductions.

Références bibliographiques

Anton Pérez, L., Gonçalo Oliveira, H., et Gomes, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, 703–717. APPIA.

Archer, V. (2009). *Graphes linguistiques multiniveau pour l'extraction de connaissances : l'exemple des collocations*. These, Université Joseph-Fourier - Grenoble I.

Assouline, P. (2007). Wikipédia, l'erreur à haut débit. *L'Histoire*, 318.

Baayen, R. H., Piepenbrock, R., et Gulikers, L. (1995). The CELEX lexical data base on CD-ROM. LDC.

Bailey, T. M. et Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44:568–591.

Baroni, M., Bernardini, S., Ferraresi, A., et Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Boula De Mareuil, P., Yvon, F., D'Alessandro, C., Aubergé, V., Vaissière, J., et Amelot, A. (2000). A French Phonetic Lexicon with variants for Speech and Language Processing. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, 273–276.

Calderone, B., Hathout, N., et Sajous, F. (2014). From GLÀFF to PsychoGLÀFF: a large psycholinguistics-oriented French lexical resource. In *Proceedings of the 16th EURALEX International Congress*, Bolzano, Italy.

Clément, L., Lang, B., et Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 1841–1844, Lisboa, Portugal.

Content, A., Mousty, P., et Radeau, M. (1990). BRULEX : Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, 90:551–566.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, 87(1):11–22.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

Dal, G. et Namer, F. (2012). Faut-il brûler les dictionnaires ? ou comment les ressources numériques ont révolutionné les recherches en morphologie. In *Actes du 3eme Congrès Mondial de Linguistique Française, Lyon*, vol. 1, 1261–1276.

Encyclopaedia Britannica (2006). Fatally Flawed: Refuting the Recent Study on Encyclopedic Accuracy by the Journal Nature.

Flaux, N., Lagae, V., et Stosic, D. (2014). Romancier, symphoniste, sculpteur : les noms d'humains créateurs d'objets idéaux. In *Actes du 4eme Congrès Mondial de Linguistique Française*, Berlin.

Giles, J. (2005). Internet Encyclopaedias go Head to Head. *Nature*, 438:900–901.

Gonçalo Oliveira, H. et Gomes, P. (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium*, 199–211. IOS Press.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., et Wirth, C. (2012). UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, 580–590.

Hathout, N., Namer, F., Plénat, M., et Tanguy, L. (2009a). La collecte et l'utilisation des données en morphologie. In Fradin, B., Kerleroux, F., et Plénat, M. (eds), *Aperçus de morphologie du français*, 267–287. Presses universitaires de Vincennes, Saint-Denis.

Hathout, N., Sajous, F., et Tanguy, L. (2009b). Looking for French deverbal nouns in an evolving Web (a short history of WAC). In *Proceedings of WAC5: Fifth Workshop on Web As Corpus*, 37–44, San-Sebastian Espagne.

Ide, N. et Véronis, J. (1994). MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 15th conference on Computational linguistics (COLING94)*, 588–592, Kyoto, Japan.

Kilgarriff, A. et Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

Martinet, A. et Walter, H. (1973). *Dictionnaire de la Prononciation Française dans son Usage Réel*. France Expansion.

Meyer, C. M. et Gurevych, I. (2012a). OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In Pazienza, M. T. et Stellato, A. (eds), *Semi-Automatic Ontology Development: Processes and Resources*, chapter 6, 131–161. IGI Global, Hershey, PA, USA.

Meyer, C. M. et Gurevych, I. (2012b). Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In Granger, S. et Paquot, M. (eds), *Electronic Lexicography*, chapter 13, 259–291. Oxford University Press, Oxford.

Mihatsch, W. et Schnedecker, C. (eds) (à paraître). *Les noms d'humains : une catégorie à part ?* Franz Steiner Verlag, Stuttgart.

Namer, F. (2003). WaliM : valider les unités morphologiquement complexes par le web. In Fradin, B., Dal, G., Kerleroux, F., Hathout, N., Plénat, M., et Roché, M. (eds), *Les unités morphologiques. Actes du 3ème Forum de Morphologie.*, 142–150, Lille.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., et Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, 19–27, Suntec, Singapore. Association for Computational Linguistics.

Navigli, R. et Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'2010)*, 216–225.

New, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. In *Verbum ex machina. Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2006)*, Louvain-la-Neuve.

Penta, D. J. (2011). The wiki-fication of the dictionary: defining lexicography in the digital age. In *Proceedings of the MiT7 Conference "unstable platforms: the promise and peril of transition"*, Cambridge, Massachusetts.

Pérennou, G. et de Calmès, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. In *Proceedings of the European Conference on Speech Technology, ECST 1987*, 1393–1396, Edinburgh, Scotland, UK.

Rajman, M., Lecomte, J., et Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

Romary, L., Salmon-Alt, S., et Francopoulo, G. (2004). Standards going concrete : from LMF to Morphalou. In Zock, M. et Saint-Dizier, P. (eds), *COLING 2004 Enhancing and using electronic dictionaries*, 22–28, Geneva. COLING.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *Proceedings of the 1st Workshop on Recent*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

Advances in Slavonic Natural Language Processing, 65–70, Brno.

Sagot, B. et Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de la 15e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2008)*, Avignon.

Sajous, F., Hathout, N., et 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)*, 285–298, Les Sables d'Olonne, France.

Sajous, F., Navarro, E., et Gaume, B. (2011). Enrichissement de lexiques sémantiques approvisionnés par les foules : le système WISIGOTH appliqué à Wiktionary. *TAL*, 52(1):11–35.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., et Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rögnvaldsson, E., et Helgadóttir, S. (eds), *Advances in Natural Language Processing*, vol. 6233 of *LNCS*, 332–344. Springer Berlin / Heidelberg.

Schlippe, T., Ochs, S., et Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH'2010)*, 2290–2293, Makuhari, Chiba, Japan.

Schultz, T., Vu, N., et Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 8126–8130, Vancouver, Canada.

Silberztein, M. (1990). Le dictionnaire électronique des mots composés. *Langue française*, 87(1):71–83.

Storkel, H. L. et Hoover, J. R. (2011). The influence of part-word phonotactic probability/neighborhood density on word learning by preschool children varying in expressive vocabulary. *Journal of Child Language*, 38:628–643.

Sérasset, G. (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul.

Urieli, A. et Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 188–201, Les Sables d'Olonne.

Zesch, T. et Gurevych, I. (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering.*, 16(01):25–59.

Zesch, T., Müller, C., et Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

---

[1] `http://fr.wiktionary.org`
Wiktionnaire désigne l'édition française de *Wiktionary*. Ce dernier désigne à la fois l'ensemble du projet et l'édition de langue anglaise.

[2] ABU : la Bibliothèque Universelle. `http://abu.cnam.fr`

[3] Lefff sert notamment à la mise au point d'analyseurs syntaxiques basés sur la théorie LFG et Morphalou est la version XML du lexique TLFnome, issu de la nomenclature du TLF.

[4] Les transcriptions phonémiques sont codées au moyen de caractères ASCII, en SAMPA ou dans un format similaire.

[5] De notre point de vue, ces ressources ne sont pas adaptées à une utilisation dans le cadre de travaux de recherche : outre leur prix élevé, le fait de ne pouvoir en redistribuer des « œuvres dérivées » constitue une limite à la portée des travaux qui les utilisent, interdisant notamment la reproductibilité des expériences. Ces ressources présentent un autre inconvénient : elles se figent une fois entrées au catalogue, faute d'être maintenues.

[6] Traduction de *crowdsourcing*. Nous limiterons notre propos aux *wikis* : *Mechanical Turk* et les « jeux à objectifs » (*games with a purpose*) entrent quant à eux dans la catégorie *Human-based computation*, que nous ne tenterons pas de traduire.

[7] `http://www.urbandictionary.com/`
Ce dictionnaire anglais propose aux internautes de donner des définitions de formes nouvelles ou existantes. Il peut s'agir notamment de formes argotiques ou de termes de sous-culture, pris ici dans son acception non péjorative, comme traduction imparfaite de l'anglais *subculture* (parfois conservé en sociologie).

[8] Nous donnons ici une description à gros grain du Wiktionnaire, en dégageant les caractéristiques principales jugées pertinentes dans la perspective de l'exploitation que nous décrivons dans les sections suivantes. En attendant une étude dictionnairique plus avancée, le lecteur intéressé pourra se référer aux travaux de Navarro et al. (2009) et Sajous et al. (2011) pour d'autres éléments de description des éditions française et anglaise, Meyer et Gurevych (2012b) pour les éditions anglaise, allemande et russe.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801216

[9] http://fr.wiktionary.org/wiki/Wiktionnaire:Critères_d'acceptation_des_articles
Nous n'avons pas corrigé les passages reproduits. On retrouve dans ces critères d'inclusion des choix opérés par d'autres lexicographes. Les formes désuètes ou archaïques encore attestées ont donc autant leur place que les créations « d'avant garde ».

[10] http://fr.wiktionary.org/wiki/Aide:Mots_apparentés
On peut trouver un exemple du flou des consignes éditoriales en consultant la page dédiée à la notion de « mots apparentés », qui, au lieu de la définir, décrit l'usage qu'en font les contributeurs.

[11] http://www.omegawiki.org/

[12] Nous présentons ici une version actualisée et enrichie de GLÀFF, par rapport à celle présentée dans (Sajous et al., 2013). Cette version a bénéficié de nombreuses corrections, de l'ajout de transcriptions phonémiques au format SAMPA, ainsi que des fréquences des lemmes et des formes dans différents corpus. Cette version est également accompagnée de l'interface d'interrogation GLÀFFOLI.

[13] http://redac.univ-tlse2.fr/lexiques/glaff.html

[14] Par exemple, l'alignement avec d'autres ressources est permis par l'ancrage des relations sémantiques au niveau des lexèmes. Dans les éditions française et anglaise, les relations relient seulement des formes graphiques et non un sens particulier de ces formes.

[15] Wiktionary au format XML : http://www.redac.univ-tlse2.fr/lexiques/wiktionaryx.html

[16] http://dumps.wikimedia.org/
Le *dump* utilisé pour ce travail est celui du 27/08/2012.

[17] Voir http://fr.wiktionary.org/wiki/Annexe:Conjugaison_en_français/marcher

[18] Pour le dire autrement, les comparaisons ne portent pas sur les 24 270 formes fléchies (resp. 13 466 lemmes) mentionnés dans les colonnes *« non simples »* du tableau 1.

[19] http://www.frantext.fr/

[20] Version du 18 juin 2008 disponible à l'adresse : http://redac.univ-tlse2.fr/corpus/wikipedia.html

[21] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[22] http://www.lexique.org/outils/Manuel_Lexique.htm

[23] Le passage de 51% d'articles mentionnant une transcription en 2010 à 90% en 2013 peut s'expliquer soit par un ajout massif de transcriptions au Wiktionnaire, soit par une meilleure détection des transcriptions par notre extracteur que par celui de Schlippe et al.

[24] Le lecteur intéressé trouvera une description plus approfondie de PsychoGLÀFF dans (Calderone et al., 2014).

[25] http://fr.wiktionary.org/wiki/Catégorie:Lexique_en_français_de_la_linguistique

[26] Voir Mihatsch et Schnedecker (à paraître), ainsi que le site du projet : http://nomsdhumains.weebly.com.

[27] http://www.crisco.unicaen.fr/des/

[28] http://fr.wiktionary.org/wiki/wikipédiholisme
Le terme correspondant (*wiktionaryholisme* ?) semble n'avoir pas encore été forgé pour Wiktionary.