SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

Construction d'un lexique flexionnel phonétisé libre du français

Olivier Bonami

Université Paris-Sorbonne, Institut Universitaire de France,
Laboratoire de Linguistique Formelle (UMR 7110, U. Paris Diderot & CNRS)
olivier.bonami@paris-sorbonne.fr

Gauthier Caron

Université de la Réunion, LCF-LIL (EA 4549)
caron.gauth@gmail.com

Clément Plancq

Laboratoire de Linguistique Formelle (UMR 7110, U. Paris Diderot & CNRS)
clement.plancq@linguist.univ-paris-diderot.fr

1 Introduction

Cet article décrit la ressource *Flexique*, un lexique flexionnel phonétisé du français standard, distribué sous licence libre (http://www.llf.cnrs.fr/flexique-fr.php). La construction de *Flexique* a été motivée par les besoins de l'étude quantitative du système flexionnel du français, et le constat d'un manque dans l'ensemble des ressources disponibles. On dispose aujourd'hui de lexiques flexionnels libres de grande qualité tels que le *Lefff* (Sagot, 2010) ou *Morphalou* (Romary, Salmon-Alt et Francopoulo, 2004), mais ceux-ci ne comportent pas de transcription phonétique des formes. De même il existe des lexiques phonétisés de grande qualité, tels que DELAP (Laporte, 1990), BDLEX (de Calmès et Pérennou, 1998) ou MHATLex (Pérennou et de Calmès, 2000), mais ceux-ci sont soit non distribués (dans le cas de DELAP), soit distribués sous des licences restrictives et à des tarifs incompatibles avec les moyens de la plupart des utilisateurs potentiels. Cette situation conduit de nombreux chercheurs travaillant sur le français, comme par exemple Bonami et Boyé (2003) ou Stump et Finkel (2013), à se limiter à l'étude d'un petit nombre de lexèmes exemplaires, avec au moins trois inconvénients majeurs : la représentativité des lexèmes exemplaires ne peut être évaluée, la fréquence des phénomènes ne peut être quantifiée, et les propriétés phonotactiques fines du lexique ne peuvent être prises en compte (Bonami et Boyé, sous presse).

La ressource *Lexique* (New, Pallier, Ferrand et Matos, 2001) comble en partie ce manque ; toutefois, pour des raisons qui sont détaillées dans la section 2, les arbitrages qui ont été faits pour le choix d'une transcription phonologique, s'ils sont adéquats pour les utilisations initialement prévues pour *Lexique* (la construction de stimuli expérimentaux pour l'étude psycholinguistique du lexique), posent problème pour l'étude linguistique du système morphologique.

Plutôt que de partir de zéro, nous avons tiré parti de la licence libre sous laquelle est distribuée *Lexique* pour construire une ressource dérivée, en adoptant des conventions explicites de transcriptions (décrites ci-après au §2.3), et en combinant la correction manuelle d'environ 65 000 formes clé, l'inférence automatique d'environ 300 000 formes supplémentaires à l'aide de fléchisseurs par règles, et la validation semi-automatique des résultats par examen des propriétés structurelles du lexique résultant.

Le présent article est structuré comme suit. Dans la section 2 nous décrivons les principales caractéristiques de la ressource. La section 3 détaille la manière dont la ressource a été construite. Dans la section 4 nous montrons quelques exemples préliminaires d'utilisation de *Flexique* pour l'étude quantitative du système flexionnel du français.

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801223

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

2 Caractéristiques de *Flexique*

Flexique a été conçu comme un outil destiné à l'étude de la structure du système flexionnel du français. Dans sa forme actuelle, cette ressource comporte trois tables correspondant respectivement aux noms, adjectifs et verbes du français, pour un total de près de 50 000 lexèmes et plus de 350 000 formes fléchies. Flexique est distribué sous licence libre[1].

catégorie	nb de lexèmes	nb de mots
nom	31 002	65 111
adjectif	11 252	45 008
verbe	4 987	253 174
Total	47 241	363 293

Tableau 1 : taille de *Flexique*

Flexique est dérivé de *Lexique* version 3.70 (New *et al.* 2001), une base de données réunissant des informations phonétiques, lexicales, morphosyntaxiques et fréquentielles sur 142 694 mots du français, et distribuée sous licence libre. *Lexique* est une ressource extrêmement utile mais peut s'avérer frustrante pour des recherches sur la flexion, pour un certain nombre de raisons :

- *Lexique* ne recense que les mots-formes attestés soit dans un sous-ensemble de textes de *Frantext* parus après 1950, soit dans le *French Subtitles Corpus* (New et Spinelli, 2013). Ainsi, les informations disponibles sur les paradigmes flexionnels sont loin d'être exhaustives; il y a notamment très peu de verbes dont le paradigme complet a été relevé.
- Parce que *Lexique* est centré sur les mots plutôt que sur les lexèmes, les formes d'un même lexème peuvent parfois ne pas avoir été décrites de manière cohérente.
- Les transcriptions phonétiques de *Lexique* sont un peu trop sommaires, pour plusieurs raisons. Il n'y a en particulier aucune représentation explicite du schwa optionnel ou de la neutralisation des voyelles médianes.
- Bien que *Lexique* fasse l'objet d'une amélioration constante, il n'a jamais fait l'objet d'une vérification manuelle minutieuse. Par conséquent, de nombreuses erreurs restent présentes par endroit, tant dans les transcriptions que dans les annotations morphosyntaxiques.

Flexique a été conçu dans le but de compléter *Lexique* dans ces domaines. En particulier, *Flexique* est organisé en lexèmes plutôt qu'en mots ; il fournit les paradigmes complets de chaque adjectif, nom et verbe dont l'une des formes au moins a été répertoriée dans *Lexique*. En outre, les transcriptions phonétiques ont été conçues pour équilibrer la fidelité à la surface et la prise en compte de la diversité des réalisations. L'idée est d'avoir pour chaque mot une représentation phonologique unique à partir de laquelle toutes les variantes phonétiques prédictibles d'un mot peuvent être déduites. Cela implique d'avoir une information systématique sur la possibilité de la présence de schwas, même lorsque ceux-ci ne sont réalisés que rarement. Cela implique également de proposer des notations spécifiques pour les segments neutralisés.

2.1 Description de la ressource

2.1.1 Format de fichier

Flexique est distribué sous la forme d'une collection de fichiers csv encodés en Unicode (UTF-8). Hormis la première ligne qui énumère les étiquettes, chacune des lignes du fichier fournit :

- un identifiant de lexème unique dérivé de l'orthographe de la forme de citation ;

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

- une liste des variantes orthographiques de la forme de citation ;
- Dans le cas des noms, une information sur le genre: « m » pour le masculin, « f » pour le féminin. Le symbole « b » (« both ») est utilisé pour les noms présentant une forme identique au masculin et au féminin (par exemple *secrétaire* /səkʁEtɛʁ/).
- Une liste des formes fléchies transcrites en quasi-API; voir section 2.3 ci-après.

Les formes inexistantes dans le paradigme des lexèmes défectifs sont indiquées par la séquence « #DEF# ».

2.2 Contenu

En l'état actuel, *Flexique* ne prend pas en compte la surabondance, c'est-à-dire les situations où plus d'une forme peut être utilisée dans une même case du paradigme (Thornton, 2012). Dans les cas de surabondance (par exemple au présent du verbe HAÏR), nous avons pris la décision de ne retenir que le patron de flexion le plus fréquent. Ainsi, chaque lexème est décrit sur une ligne, qui elle-même ne contiendra qu'une forme par colonne.

Une caractéristique notable de *Lexique* est le fait d'associer les paires de noms humains masculins et féminins à un même lemme ; par exemple *directeur* et *directrice* sont traités comme deux formes d'un même lexème. Si ce choix est celui de la tradition lexicographique française, les morphologues supposent généralement qu'en français comme dans les langues en général, les noms ont un genre unique et fixe (voir Corbett, 1991 pour le point général ; Roché, 1997 sur le genre des noms en français ; Bonami et Boyé 2005 sur la question de la dérivation productive de noms des deux genres). En même temps, il existe un nombre considérable et croissant de noms d'humains (1594 dans *Lexique*) qui s'emploient sous la même forme au masculin et au féminin (*secrétaire*, *artiste*, *prof*, etc.), et pour lesquels la postulation de deux entrées semble coûteuse. Les choix faits dans *Flexique* sont donc les suivants :

- on a codé les paires telles que *directeur/directrice* en deux entrées distinctes;
- on a utilisé une entrée unique pour les paires homophones du masculin et du féminin qui sont sémantiquement équivalentes sauf pour ce qui concerne le sexe du référent (du type *secrétaire/secrétaire*), en notant « b » (*both*) pour le genre ;
- on a utilisé deux entrées distinctes pour les paires homophones masculin-féminin n'ayant pas d'équivalence sémantique (du type *pendule*).

L'avantage de cette convention est que le lexique peut être automatiquement réajusté pour traiter les cas tels que *secrétaire* comme un ou deux lexèmes.

La délicate question de l'identité lexématique (Fradin et Kerleroux, 2009) a été traitée de manière brutale, faute de critères opératoires à grande échelle pour distinguer homophonie et polysémie. En l'état actuel, *Flexique* n'utilise qu'une ligne par paradigme (phonologique) flexionnel; Ainsi, il ne reconnaît pas de différence entre le cas d'un même lexème ayant une variante orthographique (e.g. *shaman* vs. *chamane*) et celui d'une véritable paire d'homophones (par exemple *verre* vs. *vers*). La seule exception est le cas des noms homophones ayant deux genres différents : comme nous l'avons précisé plus haut, il y a deux entrées séparées pour *pendule*. Il s'agit clairement d'une faiblesse qui pourrait facilement être évitée en couplant *Flexique* à un lexique orthographique, par exemple au *Lefff* (Sagot, 2010). Ce couplage est prévu pour la prochaine version de *Flexique*.

2.3 Conventions de transcription

Les transcriptions phonétiques utilisent les symboles API suivants:

p t k b d g f s ʃ v z ʒ m n ɲ ŋ l ʁ w ɥ j i y u e ø o ɛ œ ɔ a ə ɛ̃ ɑ̃ ɔ̃

Trois symboles ne provenant pas de l'API sont utilisés pour transcrire les voyelles médianes neutralisées:

- « E » pour la voyelle neutre entre [e] et [ɛ]

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

- « O » pour la voyelle neutre entre [o] et [ɔ]
- « Ø » pour la voyelle neutre entre [ø] et [œ]

Les symboles de l'API ont leur propre interprétation standard. Comme c'est l'habitude dans les études du français, « ə » transcrit une voyelle alternant entre [ø], [œ] et l'absence de réalisation.

Pour les formes ayant de multiples réalisations phonologiques régulières, un arbitrage a été effectué, de manière à avoir une transcription unique aussi proche que possible de la forme de surface, mais à partir de laquelle les autres réalisations possibles peuvent être déduites.

2.3.1 Schwas

Excepté en position finale de mot, les schwas ont été inclus systématiquement, même dans les cas où la réalisation effective d'un schwa est très peu fréquente. Par exemple, pour le futur de la troisième personne du singulier du verbe AIMER, la transcription est /Eməʁa/, bien qu'en parole spontanée à un débit normal, la réalisation [Emʁa] soit nettement plus fréquente. À l'inverse, on a noté comme un schwa, et non comme une voyelle fixe, une voyelle initiale qui ne tombe jamais en mention : ansi *relier* est transcrit /ʁəlje/, parce que le schwa peut tomber dans un contexte comme *il veut relier* [ilvØʁlje].

Cette décision est motivée par le fait qu'il est possible de prédire le nombre de réalisations possibles d'une forme à partir de la forme d'un mot contenant le nombre maximal de schwas internes, alors qu'il est impossible de prédire où les schwas seront possibles à partir d'une forme sans schwa. Par exemple la forme [kɔ̃tʁa] est ambigüe et correspond aussi bien à la deuxième ou troisième personne du singulier du futur du verbe COMPTER qu'au passé simple à la deuxième ou troisième personne du singulier du verbe CONTRER. De ce fait, si on choisissait cette forme comme transcription de référence, on ne pourrait en déduire si la réalisation d'un schwa intercalé est possible ou non. L'utilisation de la transcription /kɔ̃təʁa/ pour le futur de COMPTER évite cette difficulté : la possibilité de la chute du schwa interconsonantique est prédictible, et attendue dans ce contexte (entre deux syllabes ouvertes dont les attaques respectives forment ensemble une attaque branchante fréquente). Plus généralement, inclure tous les schwas est la seule solution permettant de donner une représentation phonologique unique à chaque mot-forme. Néanmoins, il faut être conscient du fait que cela réduit artificiellement la prévalence de l'homophonie. Pour cette raison, les utilisateurs de *Flexique* qui s'en serviraient pour produire des réalisations réalistes des mots-formes sont encouragés à appliquer un algorithme de syllabation réaliste pour supprimer certains schwas.

Les schwas en finale de mot ne sont pas inclus dans les transcriptions parce que leur distribution est entièrement dépendante du contexte phonologique de surface (voir par exemple Dell 1995) : il n'existe pas de contraste, par exemple, entre *ours* et *ourse*, qui sont tous les deux réalisés sans schwa final avant voyelle (*un ours allemand* [ɛ̃nuʁsalmɑ̃], *une ourse allemande* [ynuʁsalmɑ̃d]) ou en fin de groupe de souffle, mais typiquement suivis d'un schwa si cela évite un amas consonantique phonotactiquement problématique (*un ours blanc* [ɛ̃nuʁsəblɑ̃], une ourse blanche [ynuʁsəblɑ̃ʃ]).

2.3.2 Consonnes de liaison

Les transcriptions n'incluent pas les consonnes de liaison. Ceci est principalement dû à un manque de données: si la distribution de la liaison est de mieux en mieux connue, notamment grâce aux enquêtes telles que celles du projet *Phonologie du français contemporain* (voir notamment Mallet 2008), il reste une incertitude considérable quant à la disponibilité lexicale de la consonne de liaison pour les mots individuels. Ainsi, pour les adjectifs au masculin singulier, Morin (1992) et Bonami et Boyé (2005) notent que les locuteurs n'ont guère d'intuition sur la forme de liaison pour les adjectifs peu fréquents ; seule une étude à grande échelle de la distribution de la liaison *adjectif par adjectif* permettra de lever ces incertitudes. De même, en ce qui concerne les verbes au présent, il n'est pas clair que l'orthographe rende fidèlement la disponilibité de la liaison : ainsi celle-ci est-elle plus naturelle pour à la 3pl qu'à la 3sg (*il vient≠en vitesse* vs. *ils viennent=en vitesse*), et plus naturelle pour les verbes du troisièment groupe que

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

pour ceux des deux premiers (*ils arrivent en vitesse*, *ils finissent en vitesse*). Là encore, une étude empirique à grande échelle de la distribution lexicale de la liaison est nécessaire avant de pouvoir inclure des informations fiables dans une ressource lexicale comme *Flexique*.

2.3.3 Géminées

Pour les verbes ne relevant pas du premier groupe, lorsque le radical se termine en /ʁ/, les formes du futur et du conditionnel comportent un /ʁ/ géminé dans les variétés conservatrices ; par exemple la 3SG de MOURIR au futur peut être réalisée [muʁʁa]. Ce n'est bien-sûr pas la seule possibilité : la dégémination est très répandue ([muʁa]), et les régularisations par une voyelle épenthétique, bien que condamnées par la norme, sont assez fréquentes ([muʁəʁa], [muʁiʁa]). Flexique ne retient que la forme conservatrice et transcrit donc /muʁʁa/.

2.3.4 Semi-voyelles

La morphophonologie du français fait souvent alterner les voyelles hautes [i], [y] et [u], les semi-voyelles correspondantes [j], [ɥ] et [w], et les séquences voyelle-semi-voyelle [ij], [yɥ], [uw]. La convention de transcription dans *Flexique* est d'utiliser:

- une voyelle dès lors qu'il s'agit de la seule forme possible, par exemple *elle relie* : /ɛlʁəli/ ;
- une semi-voyelle dès lors qu'il s'agit de la seule forme possible, par exemple *elle paye* : /ɛlpɛj/ ;
- une séquence dès lors qu'il s'agit de la seule possibilité, par exemple *elle priait* : /ɛlpʁijE/ ;
- une semi-voyelle seule lorsque l'aternance entre une semi-voyelle et une séquence est attestée ; par exemple *tu liais*, qui peut être réalisé aussi bien [tyljɛ] que [tylijɛ], est transcrit /tyljE/.

Dans ce dernier cas, on a choisi de transcrire une forme de surface effectivement possible à partir de laquelle la variante peut être déduite sur la base de la seule observation de la transcription phonologique : dans tous les cas où une des réalisations possible comporte une attaque branchante dont le dernier segment est une semi-voyelle, la diérèse est possible ; à l'inverse, il existe des séquences voyelle-semi-voyelle qui ne peuvent pas être réduites à une semi-voyelle seule, par exemple *vous pillez* : [pije], mais pas *[pjɛ]

Là où la morphologie produit une semi-voyelle géminée (par exemple *nous payions*: /pEj+j+ɔ̃/) le locuteur du français standard réalise normalement une semi-voyelle simple. Les formes hypercorrectes telles que [pɛjjɔ̃] sont parfois entendues mais leur rareté nous a conduit à les ignorer pour les besoins de *Flexique* : *payions* est transcrit /pEjɔ̃/.

3 Construction de la ressource

Pour chaque catégorie de lexème, nous avons suivi la même procédure:

1. Détermination des *parties principales*, c'est à dire d'un sous-ensemble de cases du paradigme à partir desquelles le paradigme complet peut être déduit pour (presque) tous les lexèmes de cette catégorie (voir Stump et Finkel, 2013 sur la notion de partie principale, son statut méthodologique et son utilité théorique).
2. Filtrage des items qui sont improprement catégorisés dans *Lexique*.
3. Correction manuelle des transcriptions données par *Lexique* pour chacune des parties principales et ce pour tous les lexèmes, soit environ 65000 formes.
4. Génération automatique de paradigmes complets pour tous les lexèmes, à l'aide de fléchisseurs automatiques par règle.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

5. Vérification semi-automatique de la cohérence des paradigmes entiers, par l'examen des patrons d'alternance identifiables dans les paradigmes, sur la base des outils de Bonami et Boyé (sous presse).
6. Correction itérative tant pour les parties principales des lexèmes que pour les scripts de génération.
7. Réduction des homonymes à une seule et même description

L'aspect le plus original de la procédure est l'examen automatique des patrons d'alternance dans les paradigmes. L'idée centrale est que quand plusieurs formes d'un même lexème ont été transcrites à la main, il est très peu probable que les transcriptions soient erronées de la même manière. Par exemple, il pourrait arriver que le masculin singulier de l'adjectif COURT ait été transcrit comme /kuʁ/ et son féminin singulier comme /kurt/, avec une erreur de transcription locale au féminin, mais il y a peu de chance que la même erreur ait été commise au masculin et au féminin, les deux mots ayant été transcrits indépendamment l'un de l'autre. On utilise donc de manière opportuniste les outils mis au point par Bonami et Boyé (sous presse) pour étudier la structure implicative des paradigmes, qui basent leurs calculs sur une identification de patrons d'alternance. En examinant la trace du script, on repèrera un patron d'alternance [*X*ʁ~*X*rt] qui ne s'applique qu'à un lexème est qui est le signe évident d'une erreur. Les détails de la méthode utilisée et ses motivations sont brièvement expliqués dans le paragraphe 4.1.

Grâce à cette procédure, nous pouvons être relativement confiants que (presque) tous les lexèmes seront fléchis de manière cohérente dans *Flexique.* Nous entendons par là que les erreurs restantes seront presque toujours systématiques et générales pour tout un paradigme et non erratiques au sein d'un paradigme. Ces erreurs n'auront donc qu'un impact mineur sur l'étude du système flexionnel, qui est le domaine d'application visé par *Flexique*. Toutefois, il demeure certainement quelques erreurs systématiques. Nous prévoyons de les corriger progressivement dans les prochaines versions.

Les paragraphes suivants détaillent la manière dont la procédure ci-dessus a été mise en œuvre respectivement pour les adjectifs, les verbes et les noms.

3.1 Adjectifs

Les parties principales choisies pour les adjectifs sont le masculin singulier et le féminin singulier. La génération du pluriel à partir du singulier est aisée puisque les deux formes sont identiques, excepté un petit ensemble de cas bien connus, et au masculin seulement : principalement des adjectifs en *-al*.

Le masculin pluriel des adjectifs en *–al*, qui est notoirement imprédictible, a systématiquement été corrigé manuellement. Pour tous les autres adjectifs, partout où *Lexique* contient des formes correspondantes, au singulier comme au pluriel, nous avons vérifié si les pluriel générés concordaient avec ceux trouvés dans *Lexique*.

Il existe deux restrictions connues dans la base de données des adjectifs. La première est que nous n'avons pas généré de formes de liaison au masculin singulier, bien qu'elles ne soient pas directement déductibles à partir du reste du paradigme, et puissent être considérées comme occupant une case séparée du paradigme (voir par exemple Morin, 1992).

La seconde est que quelques adjectifs sont défectifs dans l'un des deux genres, par exemple *enceinte*. Cela n'a pas été pris en compte, et tous les adjectifs sont traités comme ayant un paradigme complet.

3.2 Verbes

La partie verbale de la ressource a été construite en trois étapes.

1. Sauf pour une poignée d'exceptions, les verbes du premier et du deuxième groupe ont un paradigme entièrement prédictible à partir de la transcription phonétique de l'infinitif et de sa

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

forme orthographique. L'utilisation de la forme orthographique est cruciale pour la désambiguation dans deux cas :

- les verbes tels que GRILLER /gʁije/, avec une semi-voyelle en finale de radical (première personne du singulier au présent *grille*: /gʁij/), ont un infinitif phonologiquement indistinguable de celui des verbes tels que CRIER /kʁije/, dont dont la semi-voyelle finale de radical alterne avec une voyelle (première personne du singulier au présent *crie*: : /kʁi/) ;
- le verbe DEJEUNER /deʒəne/, qui ne présente pas d'alterance vocalique dans la syllabe finale du radical (première personne du singulier au présent *déjeune*: /deʒən/), a un infinitif phonologiquement indistinguable de celui des verbes tels que DEMENER /dəməne/, dont le schwa alterne avec /ɛ/ (première personne du singulier au présent *démène*: /demɛn/)[2].

De ce fait, pour les verbes des deux premiers groupes, seule la transcription phonologique de l'infinitif a dû être corrigée à la main, le reste du paradigme ayant été directement généré.

2. Les paradigmes des verbes du troisième groupe ont en revanche un fort taux d'imprédictibilité qui rend l'utilisation d'une forme unique pour générer l'ensemble du paradigme impossible. Toutefois, selon Bonami et Boyé (2003), si on met de côté une poignée de verbes profondément irréguliers, le paradigme des verbes du français se décompose en 12 zones de parfaite interprédictibilité. Nous nous sommes servis de cette propriété pour sélectionner 12 parties principales, une dans chaque zone d'interprédictibilité. De ce fait, seule la transcription des douze parties principales a dû être corrigée à la main, et les 39 cases restantes ont été générées automatiquement. Concernant les quelques verbes profondément irréguliers, leurs formes exceptionnelles ont été directement incluses dans le script de génération.

3. Les verbes défectifs ont été ajoutés en dernière étape: puisque la liste des cases manquantes ne peut pas être prédite, et que la liste des verbes concernés est assez courte, les paradigmes complets ont été générés en utilisant les procédures définies ci-dessus, et la liste des formes défectives a été établie manuellement pour chaque verbe concerné.

   Il subsiste une relative incertitude quant à la liste exacte des lexèmes défectifs, et pour chacun d'entre eux, l'ensemble des cases défectives. Nous avons arbitré en tenant compte des informations réunies dans différentes publications (Boyé (2000), une édition récente du *Bescherelle* (Arrivé 1997), la dernière édition du *Bon usage* (Grevisse et 2008), et enfin le *Trésor de la Langue Française*) ainsi que des données de fréquence enregistrées dans *Lexique* et de l'usage observable sur la Toile.

3.3 Noms

Les noms n'ont pas de partie principale fiable : le paradigme d'un nom n'a que deux cases (singulier et pluriel), et on ne peut aucunement prédire une forme à partir de l'autre. Cependant le très grand nombre de noms (près de 35,000 lexèmes dans *Lexique*) rend la correction manuelle des paradigmes complets relativement laborieuse.

Pour cette raison, seules les transcriptions du singulier de *Lexique* ont été vérifiées à la main. Les pluriels ont été uniformément générés à l'identique du singulier, mais un certain nombre de vérifications ont toutefois été faites :

- Tous les noms avec une orthographe au singulier en *-al* ou *-ail* ont fait l'objet d'une vérification manuelle.
- Tous les noms n'apparaissant qu'au pluriel dans *Lexique* sont candidats à constituer des *pluralia tantum*, autrement dit des noms défectifs au singulier. Ils ont tous été vérifiés manuellement.
- Pour tous les noms dont les deux formes sont répertoriées dans *Lexique*, il a été vérifié que *Lexique* donne bel et bien un pluriel et un singulier identique, tant pour l'orthographe que

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

pour la transcription phonétique. Pour tous les lexèmes qui font exception à cette généralisation, la transcription du pluriel a été corrigée à la main.

Il demeure possible qu'un petit nombre de formes irrégulières nous ait échappé, mais celui-ci devrait s'avérer très faible, puisqu'il existe une petite quantité de noms irréguliers suffisamment inusités pour ne pas être répertoriés sous leurs deux formes (singulier et pluriel) dans *Lexique*.

C'est pour la transcription des noms que *Flexique* est le plus susceptible de comporter des inexactitudes: dans le cas des verbes et des adjectifs, plusieurs formes de chaque lexème ont été transcrites manuellement, si bien que la plupart des erreurs ont pu être relevées par l'étude des relations implicatives (elles donnent lieu à une imprédictibilité inattendue). Dans le cas des noms toutefois, cette méthode ne peut être appliquée dans la mesure où une seule forme de chaque nom a été transcrite manuellement. Les utilisateurs devront donc procéder avec précaution s'ils sont fortement dépendants de la transcription des noms.

4 Quelques exemples d'utilisation

Dans cette section nous fournissons quelques exemples illustratifs des gains descriptifs pour l'étude du système flexionnel du français permis par la ressource *Flexique*.

4.1 L'étude automatique de l'interprédictibilité dans les paradigmes de flexion

On s'appuie ici sur un domaine de recherche émergent, initié par Ackerman, Blevins et Malouf (2009), qui vise à donner une formulation quantitative à une vision strictement *Mot et paradigme* (Blevins, 2006) de la morphologie, à travers l'utilisation d'outils issus de la théorie de l'information. La démarche peut facilement être illustrée avec un exemple. Examinons, pour les adjectifs du français, la relation entre le masculin singulier et le masculin pluriel. On constate que dans l'ensemble du lexique on ne rencontre que deux patrons d'alternance : soit les deux formes sont identiques, soit le singulier est en /-al/ et le pluriel est en /-o/. De manière cruciale, l'existence de ces deux patrons d'alternance est source d'incertitude, et ce dans les deux directions : il existe des adjectifs non-alternants à singulier en /-al/, comme BANAL, et des adjectifs non-alternants à pluriel en /-o/, comme REGLO. De ce fait, un locuteur qui connaît le singulier d'un adjectif en /-al/ mais pas son pluriel se trouve dans une situation d'incertitude, alors que si le singulier ne se termine pas en /-al/, il peut inférer sans risque que le pluriel lui est identique.

On peut quantifier le niveau d'incertitude facilement, en termes d'entropie conditionnelle. Le tableau 2 donne les informations cruciales sur la distribution des patrons d'alternance dans *Flexique*.

patron	exemple		effectif
	M.SG	M.PL	
*X*al~*X*o	/mOdal/	/mOdo/	457
*X*al~*X*al	/banal/	/banal/	38
*X*o~*X*o	/ʁEglo/	/ʁEglo/	116
autre (*X*~*X*)	/pəti/	/pəti/	10 641

Tableau 2 : Distribution des patrons d'alternance entre masculin singulier et masculin pluriel des adjectifs dans *Flexique*

À partir de ces informations, on peut estimer la probabilité conditionnelle de chaque patron d'alternance sachant l'aspect phonologique d'une des formes. Ainsi, pour un adjectif dont le M.SG se termine en /-al/, la probabilité conditionnelle d'avoir un m.pl en /-o/ peut être estimée à $457/(457+38) \approx 0{,}92$. On peut également estimer la probabilité pour une case du paradigme d'être remplie par une forme ayant un certain aspect phonologique : par exemple la probabilité pour un adjectif d'avoir un masculin singulier se

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

terminant en /-al/ peut être estimée à (457+38)/11252 ≈ 0,04. Le tableau 3 précise les estimations fréquentistes des différentes probabilités pertinentes qui peuvent être déduites du tableau 2.

P(M.SG)	P(M.PL)	P(patron|M.SG)	P(patron|M.PL)
P(M.SG=*X*al) = 0,04	P(M.PL=*X*o) = 0,05	P(*X*al~*X*o | M.SG=*X*al) = 0,92	P(*X*al~*X*o | M.PL=*X*o) = 0,80
P(M.SG ≠*X*al) = 0,96	P(M.PL≠*X*o) = 0,95	P(*X*~*X* | M.SG =*X*al) = 0,08	P(*X*~*X* | M.PL=*X*o) = 0,20
		P(*X*~*X* | M.SG ≠*X*al) = 1	P(*X*~*X* | M.SG ≠*X*o) = 1

Tableau 3 : Probabilités des finales de formes et probabilités conditionnelles des patrons connaissant les finales de formes, pour les adjectifs du français, estimées à partir du tableau 2.

À partir des données du tableau 3 on peut produire une mesure globale de la prédictibilité du masculin pluriel à partir du singulier en termes d'*entropie conditionnelle*. Intuitivement, l'entropie conditionnelle du masculin pluriel connaissant le masculin singulier quantifie la diminution d'incertitude concernant la forme du masculin pluriel produite par la connaissance de la forme du masculin singulier. Dans notre cas patriculier, on arrive au résultat suivant :

$$H(\text{SG}\sim\text{PL}\,|\,\text{SG}) = -\sum_{f\in \text{SG}} P(f) \sum_{p\in \text{SG}\times PL} P(p\,|\,f)\log_2 P(p\,|\,f) =$$

$$-\left(0,04\times(0,05\times\log_2 0,05+0,95\times\log_2 0,95)+0,096\times 0\right)\approx 0,017$$

Bonami et Boyé (sous presse) appliquent ce type de mesure de manière systématique aux paradigmes verbaux du français en utilisant les transcriptions de la base BDLEX (de Calmès et Pérennou, 1998). Nous avons dans un premier temps appliqué les scripts de Bonami et Boyé de manière opportuniste à une préversion de *Flexique* pour évaluer la cohérence des transcriptions. Pour les adjectifs, nous avons entièrement examiné la liste des patrons d'alternance construite par le script pour relier les formes masculines aux formes féminines. Les incohérences de transcription dans la ressource se repèrent immédiatement dans la liste comme des patrons inattendus. Par exemple, si le M.SG de l'adjectif CANELE a été transcrit comme /kanle/ et son F.SG comme /kanəle/, le patron identifié sera [*X*le~*X*əle]. Cette méthode nous a permis de corriger plusieurs dizaines d'erreurs. Pour les verbes, il n'est pas question d'examiner à la main les listes de patrons reliant les 52×51=2652 paires de cases du paradigme. En revanche, nous avons examiné les valeurs d'entropie conditionnelle à la recherche de cas où celles-ci s'éloignent de ce qui a été trouvé par Bonami et Boyé (sous presse) — en particulier à la recherche de valeurs non-nulle quand une valeur nulle est attendue, et inversement. Dans les cas suspects les patrons d'alternance ont été examinés. La ressource a ainsi été améliorée itérativement, jusqu'à ce que l'examen de l'interprédictibilité ne révèle plus d'erreurs manifestes.

4.2 L'influence de la taille du lexique étudié sur la description des systèmes flexionnels

Les descriptions de systèmes flexionnels s'appuient souvent sur un lexique de taille réduite, surtout pour les langues peu dotées, mais parfois aussi pour les langues bien décrites comme le français. Le choix d'un lexique de taille réduite peut sembler raisonnable : il est banal d'observer que les lexèmes irréguliers sont aussi les plus fréquents, si bien que l'examen d'un lexique complet est présumé ne pas faire découvrir d'observations nouvelles.

La taille de *Flexique* permet d'examiner empiriquement la validité de cette stratégie, en procédant à des expériences d'échantillonage. Pour ce faire, on a successivement constitué à partir du *French Subtitles Corpus* (New et Spinelli, 2013), 50 échantillons aléatoires de *n* adjectifs, pour *n* variant de 500 à 10000 avec un pas de 500—soit 1000 échantillons au total. On a ensuite calculé, à partir des transcriptions de *Flexique*, l'entropie conditionnelle de *B* connaissant *A*, pour chaque paire de cases du paradigme *A* et *B,*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

pour chaque échantillon. La figure 1 présente les principaux résultats : à chaque taille d'échantillon, et pour trois paires de cases du paradigme, on a indiqué l'entropie conditionnelle moyenne sur les 50 échantillons tirés, ainsi que l'écart-type.

Si une étude plus fouillée s'impose, la simple lecture de la figure donne trois résultats très nets. D'abord, il existe des patrons de flexion réellement rares. Dans la courbe concernant la prédiction du féminin singulier à partir du masculin singulier, après un plateau entre 1500 et 4500 lexèmes, on observe une nette augmentation quand la taille du lexique atteint 5000 lexèmes ; celle-ci ne peut être due qu'au fait qu'il existe des patrons qui n'avaient pas été rencontrés à des tailles d'échantillon plus petites. Ensuite, à des petites tailles d'échantillon (inférieures à 1000), non seulement l'incertitude peut être fortement sous-estimée, mais la variance est considérable. Deux linguistes de bonne foi qui baseraient chacun leur étude sur les 500 premiers adjectifs qu'ils rencontrent arriveraient donc à des résultats différents. Enfin, la stabilisation de l'entropie et la baisse de la variance interviennent à des tailles de lexique différentes en fonction des cases du paradigme examinées. Ainsi, pour la prédiction du masculin pluriel à partir du masculin singulier, l'estimation de l'entropie est excellente dès 500 lexèmes. On ne peut donc pas donner de principe méthodologique simple quant à la taille de lexique minimale pour procéder à une étude informative.

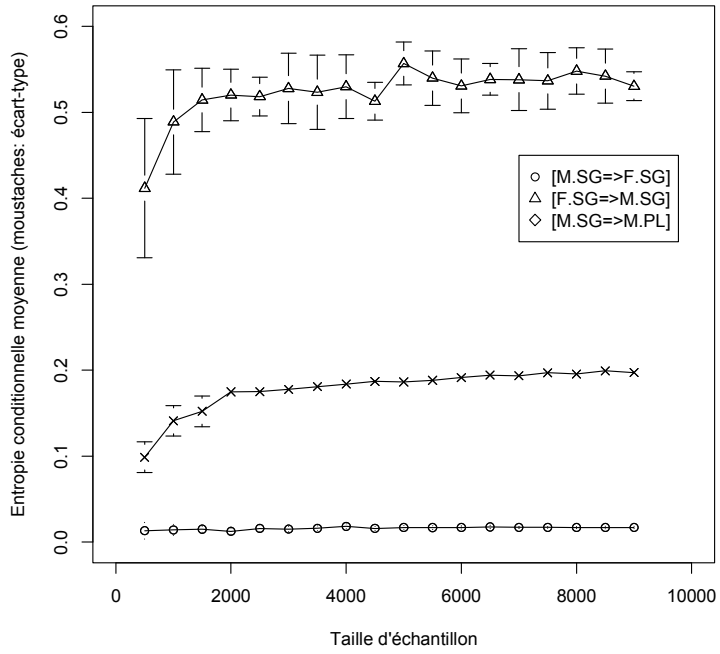

Figure 1 : Variation de l'entropie conditionelle entre cases du paradigme des adjectifs du français en fonction de la taille du lexique

4.3 L'inférence automatique de classes flexionnelles

Un des outils de base de la description grammaticale est la constitution de classes flexionnelles hiérarchiquement organisées : les lexèmes sont regroupés en classes et superclasses sur la base de la ressemblance entre les patrons de flexion qu'ils exemplifient ; Kilani-Schoch et Dressler (2005) est sans aucun doute la tentative la plus aboutie pour la conjugaison du français. Ce travail de classification est habituellement fastidieusement fait à la main, et en utilisant des combinaisons de critères de types variés et dont le poids relatif est jugé de manière impressioniste et pas forcément cohérente sur l'ensemble du lexique.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

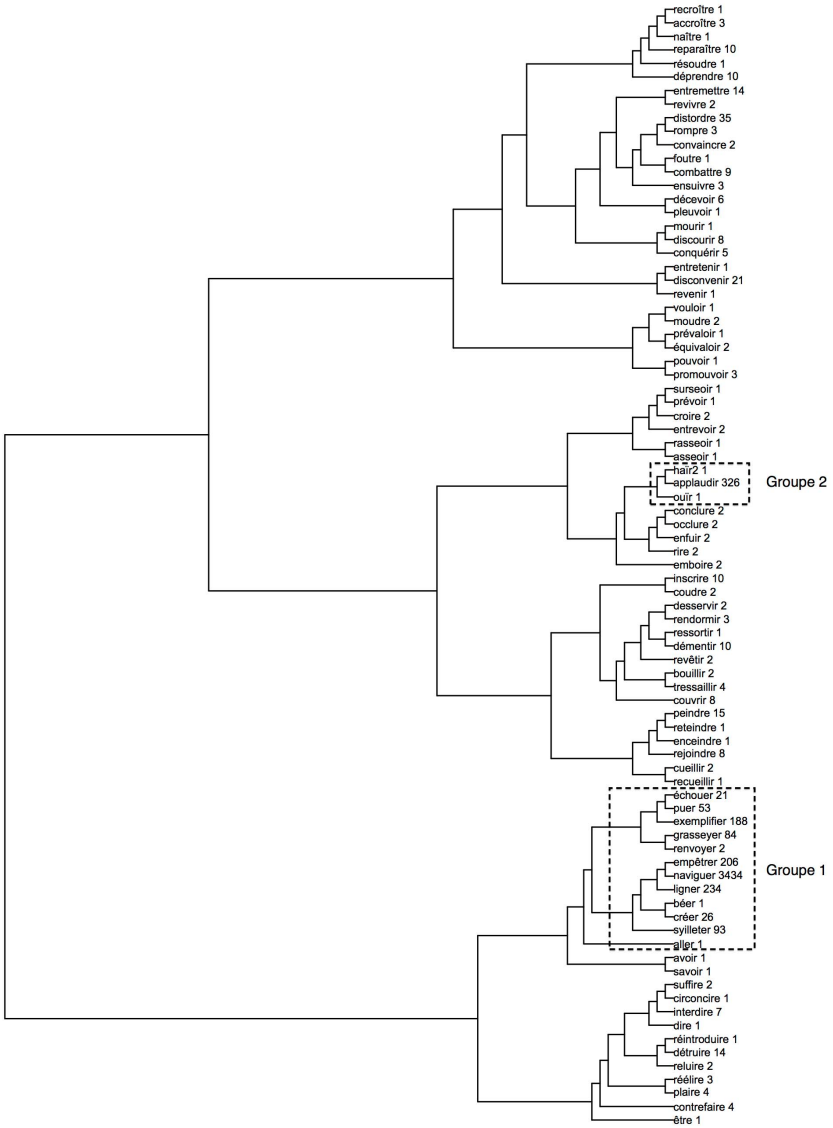

Figure 2 : Dendrogramme de la classification hiérarchique des verbes de *Flexique*.

La disponibilité de lexiques flexionnels à grande échelle comme *Flexique* permet d'envisager une méthode objectivée et automatique. Nous donnons ici l'exemple d'une possibilité parmi de nombreuses autres, qui est dans l'esprit de ce que proposent Brown et Evans (2012) tout en étant très différent dans l'exécution. Dans un premier temps, on a classé tous les verbes de *Flexique* en fonction des patrons d'alternance reliant chaque paire de cases de leurs paradigmes. Le patron flexionnel de chaque verbe est donc représenté par un long vecteur indiquant successivement, pour chaque paire de cases, le patron que ce verbe exemplifie. Dans un deuxième temps, on calcule pour chaque paire de lexèmes la distance de Hamming entre leurs patrons flexionnels—autrement dit le nombre de paires de cases pour lesquels ils utilisent des patrons distincts. Dans un troisième temps, la matrice de distances obtenue est donnée en entrée à un algorithme de classification. La figure 2 ci-dessus présente le dendrogramme obtenu pour les verbes du français en procédant à une classification hiérarchique ascendante par connexité moyenne (Sokal et Michener, 1958). Les étiquettes de feuilles indiquent un des verbes appartenant à la micro-classe de lexèmes ayant une distance 0 ainsi que l'effectif de cette micro-classe.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

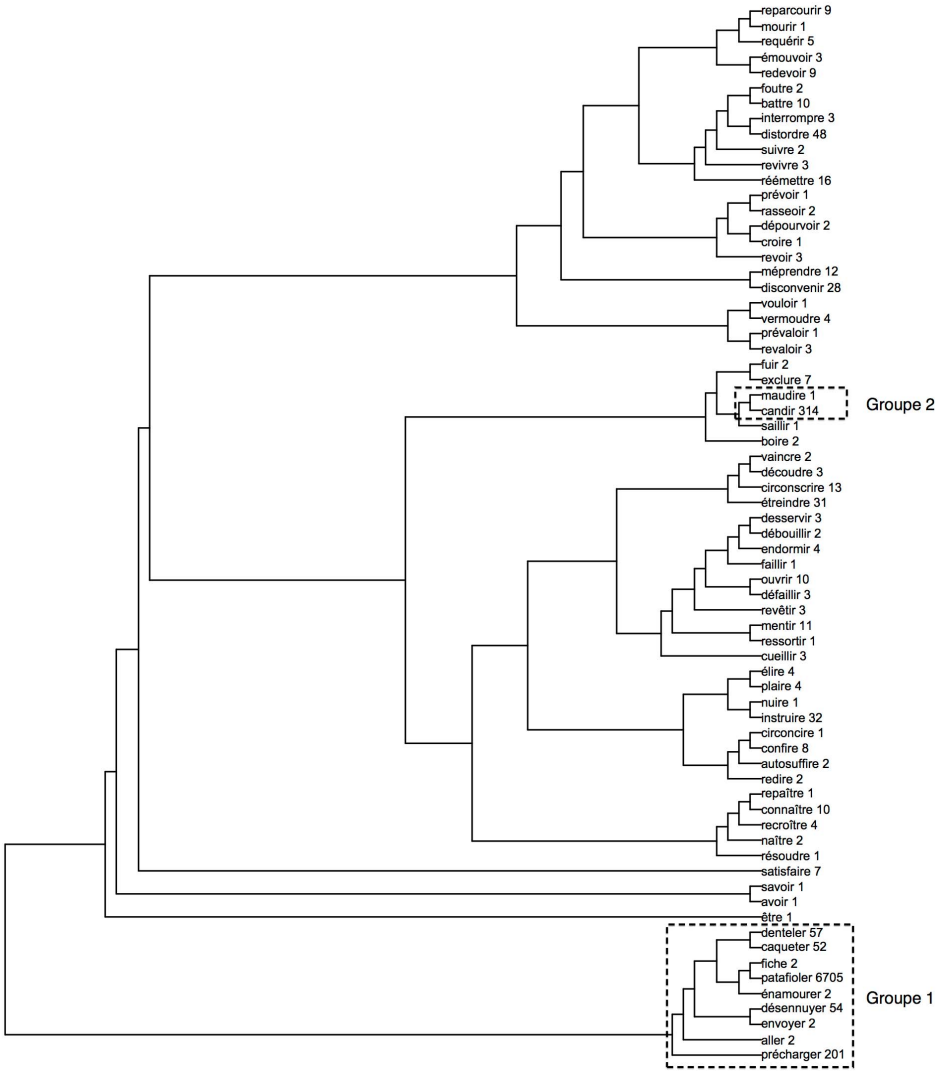

Figure 3 : Dendrogramme de la classification hiérarchique des verbes de *Flexique*.

Un examen rapide montre que les deux premiers groupes de conjugaison sont correctement identifiés, et que la sous-classification semble intuitivement correcte. Un examen plus détaillé, et une comparaison avec les classifications produites manuellement ou celles qui ont été produites par d'autres méthodes automatiques, promet d'enrichir la réflexion sur les critères conduisant au postulat d'un système de classes flexionnelles donnés pour une langue.

4.4 La flexion du français oral et la flexion du français écrit

La littérature pédagogique sur le français regorge d'observations superficielles sur les différences entre le système flexionnel oral et le système qui résulte des conventions orthographiques. Il est indéniable qu'il existe des différences nettes entre les deux systèmes, et il est généralement admis, sans évaluation concrète, que le système orthographique est plus complexe que le système oral, en particulier parce qu'il utilise parfois plusieurs stratégies pour coder un même patron d'alternance.

On peut utiliser *Flexique*, en combinaison avec un des grands lexiques orthographiques existants, pour aborder cette question de manière objectivée. On a appliqué exactement la même méthode décrite dans le

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

paragraphe 4.3 aux 7745 verbes non-défectifs du *Lefff* (Sagot 2010). La figure 3 ci-dessus présente le dendrogramme obtenu. Deux observations massives s'imposent. D'une part, le nombre de micro-classes est plus petit pour le français écrit que pour le français oral (69 vs 83), et ce bien que le lexique écrit utilisé soit plus grand de près de 50% ; cette observation semble aller à l'encontre de l'idée selon laquelle la conjugaison écrite est plus complexe. D'autre part et surtout, la forme de la classification est nettement différente : par exemple, si les deux classifications s'accordent à identifier le premier groupe de verbes, celui-ci forme une grappe séparée pour l'écrit mais pas pour l'oral, où il se regroupe avec une partie du troisième groupe.

Références bibliographiques

Ackerman, F., Blevins, J., et Malouf, R. (2009). Parts and wholes: Implicative patterns in inflectional paradigms. Blevins, J. et Blevins, J. (éds), *Analogy in Grammar: Form and Acquisition*. Oxford : Oxford University Press, 54-82.

Arrivé (1997). *Le Bescherelle : La conjugaison pour tous.* Paris : Hatier.

Blevins, J. (2006). Word-based morphology. *Journal of Linguistics, 42,* 531-573.

Bonami, O. et Boyé, G. (2003). Supplétion et classes flexionnelles dans la conjugaison du français. *Langages, 152,* 102-126.

Bonami, O. et Boyé, G. (2005). Construire le paradigme d'un adjectif. *Recherches linguistiques de Vincennes, 34,* 77-98

Bonami, O. et Boyé, G. (Sous presse). De formes en thèmes. Villoing, F., Leroy, S. et David, S. (éds.), *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Nanterre : Presses de l'Université Paris-Ouest.

Boyé, G. (2000). *Problèmes de morpho-phonologie verbale en français, en espagnol et en italien.* Thèse de doctorat, Université Paris 7.

Brown, D. et Evans, R. (2012). Morphological complexity and unsupervised learning: validating Russian inflectional classes using high frequency data'. Kiefer, F., Ladányi, M., et Siptár, P. (eds.), *Current Issues in Morphological Theory: (Ir)regularity, analogy and frequency*. Amsterdam: John Benjamins, 135-162.

Corbett, G. (1991). *Gender*. Cambridge : Cambridge University Press.

de Calmès, M. & Pérennou, G. (1998). BDLEX: a Lexicon for Spoken and Written French. *1st International Conference on Language Resources and Evaluation*, 1129–1136.

Dell, F. (1995). Consonant clusters and phonological syllables in French. *Lingua, 95,* 5-26.

Fradin, B. et Kerleroux, F. (2009). L'identité lexémique. Fradin, B., Kerleroux, F. et Plénat, M. (eds), *Aperçus de morphologie du français*. Saint-Denis: Presses Universitaires de Vincennes, 85-104.

Grevisse, M. et Goosse, A. (2008). *Le bon usage*, 14e édition. Bruxelles: De Boeck.

Kilani-Schoch, M. et Dressler, W. (2005). *Morphologie naturelle et flexion du verbe français*. Tübingen : Gunter Narr Verlag.

Laporte, E. (1990). Le dictionnaire phonémique DELAP. *Langue française, 87*, 59-70.

Mallet, G. (2008). *La liaison en français : descriptions et analyses dans le corpus PFC*. Thèse de l'U. Paris Ouest Nanterre La Défense.

Morin, Y.-C. (1992). Un cas méconnu de la déclinaison de l'adjectif en français: les formes de liaison de l'adjectif antéposé, *Le mot, les mots, les bons mots. Word, words, witty words. Hommage à Igor A. Mel'čuk*. Montréal : Presses de l'Université de Montréal, 233-250.

New B., Pallier C., Ferrand L., Matos R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101, 447-462. http://www.lexique.org

Pérennou, G. & de Calmès, M. (2000). MHATLex: Lexical Resources for Modelling the French Pronunciation. *Proceedings of LREC 2000*.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801223

Roché, M. (1997). *La variation non flexionnelle du genre des noms. Diachronie, diatopie, diastratie*. *Cahiers d'Etudes Romanes*, hors série, Toulouse.

Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete : from LMF to Morphalou. *Workshop on Electronic Dictionaries, Coling 20*.

Sagot, B. (2010). The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French. *Proceedings of LREC 2010*.

Sokal, R. et Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38,* 1409-1438.

Stump, G. et Finkel, R. (2013). *Morphological typology: from Word to Paradigm*. Cambridge : Cambridge University Press.

Thornton, A. (2012). Reduction and maintenance of overabundance. A case study on Italian verb paradigms". *Word Structure*, *5,* 183-207.

---

[1] La licence utilisée est la même que celle de *Lexique,* à savoir une licence Creative Commons Attribution-NonCommercial-ShareAlike (http://creativecommons.org/licenses/by-nc-sa/3.0/).

[2] Le seul autre exemple de ce type est le verbe BECQUETER pour les locuteurs qui peuvent prononcer un schwa dans la deuxième syllabe : infinitif [bɛkəte] présent 3SG [bɛkət]. Cette réalisation semblant minoritaire : la plupart des locuteurs conjuguent BECQUETER soit selon le modèle de JETER, soit selon celui de JACTER. C'est cette dernière conjugaison qui est adoptée dans *Flexique*, qui transcrit donc l'infinitif /bEkte/.