Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Des ressources lexicales du français et de leur utilisation en TAL : étude des actes de TALN

Hee-Soo Choi[1, 2] Karën Fort[2, 3] Bruno Guillaume[2] Mathieu Constant[1]
(1) ATILF, CNRS, Université de Lorraine, 54000 Nancy, France
(2) LORIA, Université de Lorraine, 54506 Vandoeuvre-lès-Nancy, France
(3) Sorbonne Université, 75006 Paris, France
hee-soo.choi@loria.fr, karen.fort@loria.fr,
bruno.guillaume@loria.fr, mathieu.constant@atilf.fr

RÉSUMÉ

Au début du XXIe siècle, le français faisait encore partie des langues peu dotées. Grâce aux efforts de la communauté française du traitement automatique des langues (TAL), de nombreuses ressources librement disponibles ont été produites, dont des lexiques du français. À travers cet article, nous nous intéressons à leur devenir dans la communauté par le prisme des actes de la conférence TALN sur une période de 20 ans.

ABSTRACT

French lexicons and their usage in NLP : a study of TALN proceedings.

At the beginning of the 21st century, French was still considered a less-resourced language. Thanks to the efforts of the French Natural Language Processing (NLP) community, many freely available resources have been produced, including French lexicons. We are interested here in their future in the community through the prism of the proceedings of the TALN conference over a 20-year period.

MOTS-CLÉS : lexiques, ressources lexicales, français.

KEYWORDS: lexica, lexical resources, French.

1 Introduction

Depuis le début des années 2000, de nombreux efforts ont été réalisés pour créer des ressources langagières pour le français, notamment lexicales, qui soient librement disponibles. En 2019, Mariani *et al.* (2019b) démontrent l'importance de ces ressources en recueillant le nombre de mentions d'une large variété de ressources langagières sur une période de 50 ans, distinguant les lexiques, des corpus et des outils de Traitement Automatique des Langues (TAL). La ressource langagière la plus mentionnée est `WordNet` (Fellbaum, 1998) suivi de trois corpus `Timit`, `Wikipedia` et le `Penn Treebank` (Marcus *et al.*, 1993). Mais qu'en est-il pour les ressources lexicales du français ? Les efforts de la communauté ont-ils portés ? Inspirés des travaux de Mariani *et al.* (2019a,b), notre étude présente la dynamique des ressources lexicales du français librement disponibles dans les actes de la conférence TALN entre 2001 et 2022. À travers ce travail, nous nous intéressons à la manière dont ces ressources sont utilisées, par qui et sur quelles applications.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 23–36.

2 Les ressources lexicales publiées à TALN

2.1 Des lexiques aux plongements lexicaux statiques

Il existe aujourd'hui de nombreuses ressources lexicales du français utilisables en TAL, de différents types (lexiques, réseaux lexico-sémantiques, corpus...) et de différents niveaux de description linguistique (syntaxe, sémantique, morphologie...) (Gala, 2013). Si les premières ressources ont été créées manuellement par des linguistes, la communauté de TAL s'est efforcée de concevoir des ressources lexicales informatisées adaptées à ses besoins, notamment celui d'obtenir une ressource conséquente et la plus complète possible. Ainsi, une large variété de ressources a vu le jour grâce à des constructions semi-automatique, par myriadisation ou par fusion de ressources existantes (Choi, 2022).

Recenser une liste exhaustive de ressources constitue alors une tâche délicate dans la mesure où un certain nombre de ressources en englobent d'autres. De ce fait, afin de fixer un critère de sélection, nous avons considéré tous les lexiques du français présents dans la LRE map [1] et Ortolang [2] dont les liens étaient accessibles pour télécharger la ressource. Si nous faisons le choix de considérer les ressources multilingues, nous décidons de laisser de côté Wikipédia et Wiktionnaire qui n'ont pas été créés initialement pour la recherche ainsi que les lexiques spécialisés et de langues régionales. Nous obtenons ainsi une liste finale de 31 ressources présentées dans le tableau 1. Les personnes impliquées dans la conception des ressources correspondent aux auteurs des articles de référence et aux personnes présentées comme contributeurs dans les sites officiels des ressources.

Ressources	Création	Personnes impliquées dans la conception	Licence
Lexique-Grammaire (Gross, 1975)	Fin 1960	M. Gross et al. (LADL) [3]	LGPL-LR
DELA	1990	M. Gross et al. (LADL)	LGPL-LR
LVF (Dubois & Dubois-Charlier, 1997)	1997	J. Dubois, F. Dubois-Charlier	LGPL-LR
Dictionnaire Électronique des Synonymes (DES) (Ploux & Victorri, 1998)	1998	S. Ploux, B. Victorri, J-L. Manguin, M. Morel, L. Chardon	CC BY-NC-SA
Dicovalence (Van den Eynde & Mertens, 2006, 2010; Mertens, 2010)	2003	K. van den Eynde, P. Mertens	LGPL-LR
Lefff (Clément *et al.*, 2004; Sagot, 2010)	2003	B. Sagot, L. Clément	LGPL-LR
ProLexBase (Tran & Maurel, 2006)	2006	M. Tran, D. Maurel	LGPL-LR
Jibiki (Mangeot & Chalvin, 2006; Mangeot-Nagata, 2016)	2006	M. Mangeot-Nagata, A. Chalvin	Domaine Public
JeuxDeMots (Lafourcade & Joubert, 2008; Lafourcade & Le Brun, 2020)	2007	M. Lafourcade, A. Joubert, N. Le Brun	Domaine Public
VfrLPL (Rauzy & Blache, 2007)	2007	S. Rauzy, P. Blache	CRDO
LGLex (Constant & Tolone, 2010)	2008	E. Tolone, M. Constant	LGPL-LR
WOLF (Sagot & Fišer, 2008)	2008	B. Sagot, D. Fišer	CeCILL-C
RL-Fr (Lux-Pogodalla & Polguère, 2011)	2011	V. Lux-Pogodalla, A. Polguère, S. Ollinger	CC BY
DBnary (Sérasset, 2015)	2012	G. Sérasset	CC BY-SA
Dictionnaire morphosyntaxique du français (DM) (Trouilleux, 2012)	2012	F. Trouilleux	GNU GPL
DiLAF (Enguehard *et al.*, 2012)	2012	C. Enguehard, S. Kané, M. Mangeot, I. Modi, M. Sanogo	CC BY-SA
GLÀFF (Sajous *et al.*, 2013)	2013	F. Sajous, N. Hathout, B. Calderone	CC BY-SA
Marsalex (Blache & Rauzy, 2008)	2013	P. Blache, S. Rauzy	CC BY
Démonette (Hathout & Namer, 2014)	2014	N. Hathout, F. Namer	CC BY-NC
FLELex (François *et al.*, 2014)	2014	T. François, N. Gala, P. Watrin, C. Fairon, A. Pintard	CC BY-NC-SA
French FrameNet (Candito *et al.*, 2014)	2014	M. Candito, P. Amsili, L. Barque, F. Benamara, G. Chalendar, L. Vieu, M. Djemaa, P. Haas, R. Huyghe, Y. Mathieu, P. Muller, B. Sagot	LGPL-LR
VerbeNet (Danlos *et al.*, 2014)	2014	L. Danlos, T. Nakamura Q. Pradet	CC BY-SA
OpeNER-sentiment-lexicons (Maks *et al.*, 2014)	2014	I. Maks, R. Izquierdo, F. Frontini, R. Agerri, P. Vossen, A. Azpeitia	OpenSource
Morphalou (ATILF, 2019)	2015	S. Ollinger, C. Benzitoun, E. Jacquey, U. Fleury	LGPL-LR
TLFPhraseo (ATILF, 2016)	2016	E. Jacquey, J. Humbert	CC BY-NC-SA
Apertium RDF Graph (Villegas *et al.*, 2016)	2016	M. Villegas, M. Melero, N. Bel, J. Gracia	CC BY-SA
ReSyf (Gala *et al.*, 2013)	2018	N. Gala, M. Billami, T. François, C. Fairon, D. Bernhard	CC BY-NC
Nomage (Balvet *et al.*, 2011)	2019	A. Balvet, L. Barque, M. Condette, R. Huyghe, A. Jugnet, R. Marin, A. Merlo, P. Haas	LGPL-LR
VerNom (Missud *et al.*, 2020)	2020	A. Missud, P. Amsili, F. Villoing	CC BY-NC-SA
Holinet (Prost, 2022)	2022	J-P. Prost	CC BY
Lexique4linguists (Schalchli, 2022)	2022	G. Schalchli	CC BY

TABLE 1 – Liste des 31 lexiques sélectionnés du plus ancien au plus récent.

1. https://lremap.elra.info/?type=Lexicon&availability=Freely+Available&languages=french, consultée le 16 mai 2023.
2. https://www.ortolang.fr/market/lexicons, consultée le 16 mai 2023.

Outre les lexiques, nous observons également les occurrences de trois plongements lexicaux statiques disponibles pour le français : `Word2vec` (Mikolov *et al.*, 2013; Fauconnier, 2015), `FastText` (Bojanowski *et al.*, 2017; Grave *et al.*, 2018) et `GloVe` (Pennington *et al.*, 2014). Grâce à leurs représentations vectorielles codant des informations linguistiques, les plongements lexicaux constituent une autre forme de ressource lexicale, de plus en plus utilisés en TAL depuis leur apparition en 2013.

2.2 Corpus TALN 2001-2022

Notre corpus est composé d'articles de TALN de 2001 à 2022 extraits d'ACL Anthology au format PDF, convertis au format XML avec GROBID (Lopez, 2008 2023). Au total, 2 176 articles en PDF et 2 174 articles en XML ont été récupérés, deux articles de 2010 n'ayant pas été convertis. La conversion pouvant être défectueuse, une vérification semi-automatique a été faite en deux temps : nous avons déterminé si la balise *lang* correspond à celle du français ou de l'anglais et si des mots-outils de la langue concernée sont présents dans le corps de l'article. Nous avons ainsi recensé 80 articles mal convertis dont 39 pour l'année 2013 et 24 pour l'année 2001 [4]. Par ailleurs, nous avons fait le choix de retirer les articles JEP, invités, tutoriels, DEFT et des ateliers, ceux-ci pouvant biaiser les résultats. En effet, en 2014, deux ateliers sur les ressources ont eu lieu (FondamenTAL et RLTLN), qui ont largement augmenté les chiffres obtenus. Notre corpus final contient 1 511 articles exploitables au format XML (cf. Annexes).

Afin d'observer l'utilisation des lexiques sélectionnés dans les articles, les occurrences de chaque ressource dans le corps de l'article ont été extraites automatiquement. Nous avons décidé de considérer uniquement le corps de l'article avec la balise *body* pour éviter les biais dûs aux occurrences dans les résumés et la bibliographie. En amont, des pré-traitements simples tels qu'une suppression de certaines ponctuations, des majuscules et une tokenisation ont été appliqués. Nous avons également veillé à normaliser autant que faire se peut les noms des ressources. Par exemple, *glaff* est remplacé par *glàfff*, plus fréquent. Au total, seuls 231 articles présentent au moins un des lexiques considérés, ce qui représente environ de 15 % du corpus entier. De plus, deux lexiques n'apparaissent dans aucun article du corpus : `TLFPhraseo` et `Lexique4linguists`. Concernant les plongements lexicaux, 108 articles présentent au moins une des trois ressources entre 2014 et 2022.

En raison des difficultés à extraire des informations présentes à partir d'occurrences brutes, nous avons décidé de faire une annotation manuelle sur les articles présentant au moins un lexique ou des plongements lexicaux. Pour chaque article et pour chaque ressource présente dans celui-ci, on attribue une classe parmi les cinq suivantes :

— **Construction-Extension** : l'article présente comment la ressource a été élaborée ou traite de son extension/amélioration.
— **Utilisation** : l'article présente une expérience où la ressource est utilisée pour une tâche spécifique (construction d'une autre ressource, étiquetage morpho-syntaxique...).
— **Comparaison** : l'article utilise la ressource comme une référence pour faire une comparaison avec une autre ressource.
— **Mention** : l'article ne fait que mentionner la ressource (état de l'art).
— **Erreur** : l'article présente une fausse occurrence (à cause d'une erreur dans le fichier XML),

3. Pour le Lexique-Grammaire et DELA, les personnes considérées sont des personnes passées par le LADL ayant utilisé la ressource comme L. Danlos, E. Tolone, S. Voyatzi, T. Nakaruma, E. Laporte, S. Paumier. Cette liste n'est pas exhaustive.

4. Nous notons que le corpus ACL Anthology (Rohatgi, 2022) présente les mêmes articles défectueux. L'erreur de conversion provient vraisemblablement d'un problème dans les articles d'origine.

l'occurrence ne correspond pas au nom de la ressource (Ex : « wolf » peut désigner un nom propre), la ressource ne concerne pas le français (Ex : `FrameNet` de l'anglais, plongements lexicaux de `word2vec` ou lexiques multilingues utilisés pour une autre langue), l'article utilise l'outil et non la ressource (modèle `word2vec` et non les plongements lexicaux existants).

L'annotation a été effectuée en séparant les lexiques des plongements lexicaux. Tous les articles présentant un lexique ont été annotés par un des auteurs (annotateur 0). Un accord inter-annotateur a été calculé sur un échantillon de 20 articles avec deux paires d'annotateurs (annotateur 0 - annotateur 1, annotateur 0 - annotateur 2). Sur 20 articles, les annotateurs sont en accord sur 19 d'entre eux. Les articles contenant des plongements lexicaux ont été annotés par trois annotateurs. Un accord inter-annotateurs a été calculé de la même manière que pour les lexiques et montre que sur 20 articles, les annotateurs sont en accord sur 17 d'entre eux.

La figure 1 présente la distribution dans le temps des lexiques en fonction du type d'usage. Nous pouvons observer que les articles présentant les lexiques augmentent progressivement de 2004 à 2014, où un pic est atteint avec 27 articles. Cette année-là, plusieurs ressources voient le jour comme le `French FrameNet` (Candito *et al.*, 2014) et `Démonette` (Hathout & Namer, 2014). Cette dernière étant une fusion de ressources existantes, le nombre de ressources utilisées augmente également. D'autres ressources font également l'objet d'extension comme le `RL-Fr` (Lux-Pogodalla, 2014) ou `JeuxDeMots` (Lafourcade *et al.*, 2014). Après 2014, le nombre d'articles diminue mais se maintient autour d'une dizaine d'articles jusqu'en 2021 où l'on en compte seulement deux.

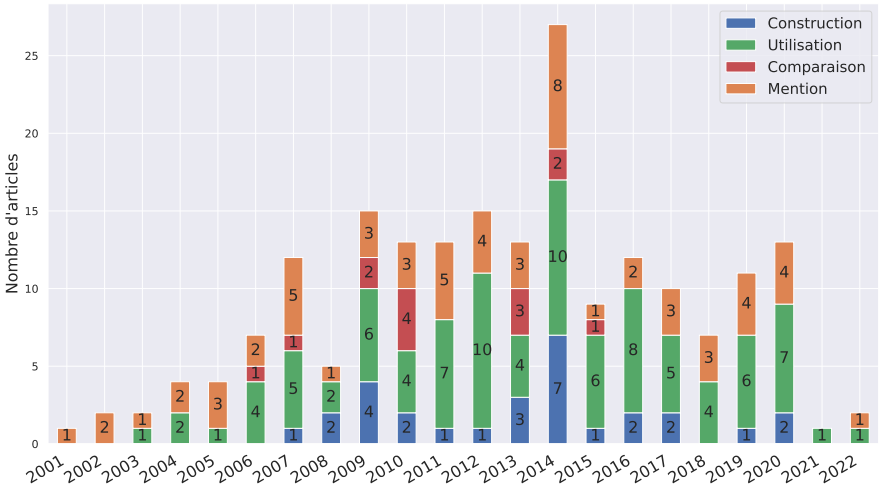

FIGURE 1 – Distribution dans le temps du nombre d'articles présentant au moins un lexique selon le type d'usage.

Concernant les plongements lexicaux, nous avons veillé à faire la distinction entre la ressource et l'outil. En effet, nous avons annoté en « Construction » le cas où les auteurs génèrent des plongements lexicaux et les mettent à disposition, contrairement au cas où le modèle est utilisé pour une application précise sans redistribution des plongements générés. Par ailleurs, comme pour les lexiques multilingues, nous ne prenons en compte que les articles traitant du français. Ils sont au nombre de 23 contre 50 pour l'anglais. Nous notons que cette vérification ne s'est pas faite de manière triviale en raison d'un certain nombre d'articles ne mentionnant pas explicitement la langue traitée (Ducel *et al.*, 2022).

3 Pour quoi et par qui sont utilisées les ressources lexicales ?

3.1 Des applications de différents types

La figure 2 présente la distribution en type d'usage pour 16 lexiques, telle qu'identifiée manuellement. Par souci de lisibilité, nous ne présentons pas les 15 lexiques restants, leurs occurrences étant strictement inférieures à 2, toutes classes confondues. Les ressources les plus présentes dans le corpus sont le Lefff (59 articles), JeuxDeMots (36 articles), Lexique-Grammaire (25 articles), Morphalou (21 articles) et WOLF (20 articles). En termes d'utilisation, le Lefff se distingue significativement des autres avec 36 articles. Morphalou et WOLF apparaissent tous deux dans une vingtaine d'articles cependant WOLF ne fait l'objet d'une utilisation que dans cinq d'entre eux tandis que Morphalou est utilisé dans dix articles. Les ressources les plus utilisées sont des lexiques présentant des contenus linguistiques différents. Si le Lefff et Morphalou sont des lexiques des formes fléchies du français, JeuxDeMots présente principalement des relations sémantiques entre les mots (voire des termes) sous la forme d'un réseau lexico-sémantique. De ce fait, ces ressources ne sont pas exploitées pour les mêmes applications.

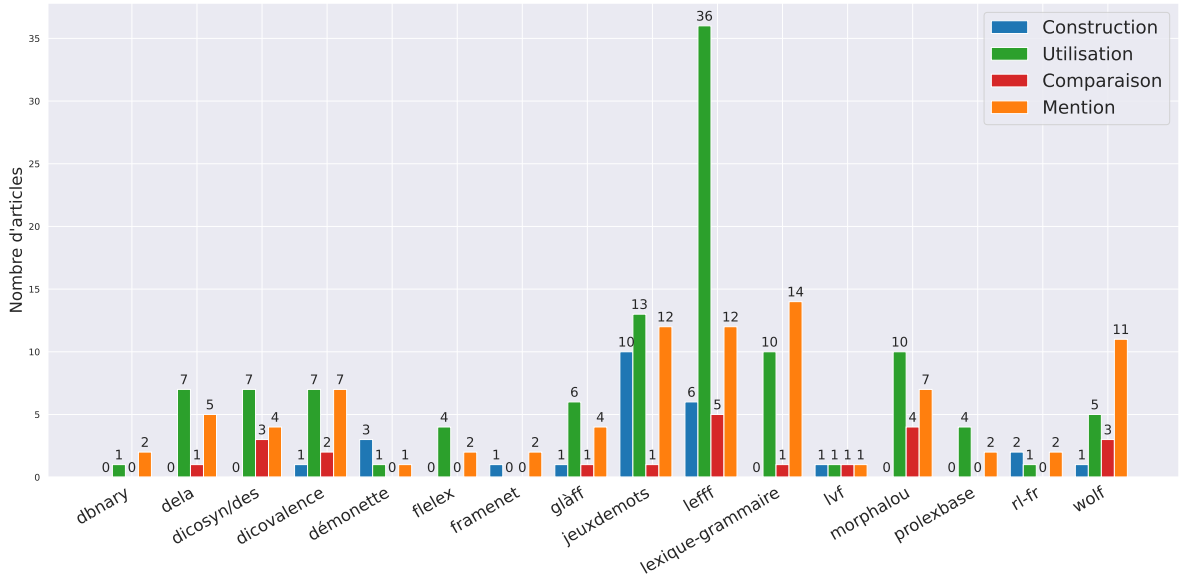

FIGURE 2 – Distribution des ressources les plus fréquentes selon le type d'usage.

Dans le domaine de la syntaxe et de la morphologie, le Lefff a été utilisé dans la construction de lexiques spécifiques (Strnadová & Sagot, 2011; Sagot, 2019), l'intégration à des étiqueteurs en parties du discours tels que MElt (Denis & Sagot, 2012), LGTagger (Constant & Sigogne, 2011) ou Macaon (Nasr *et al.*, 2011), la traduction automatique (Bawden, 2017; Burlot & Yvon, 2018), la fouille d'erreurs dans des sorties d'analyseurs syntaxique (Sagot & Villemonte De La Clergerie, 2006) ou la correction de textes bruités (Baranes, 2012).

En sémantique, JeuxDeMots est principalement exploité pour des travaux sur les relations sémantiques : l'hyperonymie et l'hyponymie (Gosset *et al.*, 2021), la synonymie (Francois *et al.*, 2016) ou la méronymie (Morlane-Hondère & Fabre, 2012). La ressource présente la particularité d'apparaître dans approximativement le même nombre d'articles qui décrivent sa construction et qui l'utilisent. Cela s'explique par les différentes extensions qu'a connues la ressource comme le jeu ColorIt permettant

de faire des associations de mots et de couleurs (Lafourcade *et al.*, 2014).

La figure 3 montre que les plongements lexicaux les plus utilisés sont `FastText` et `Word2vec`. Dans notre corpus, ils interviennent dans des tâches telles que la classification et l'extraction de relations (Khaldi *et al.*, 2020; Randriatsitohaina & Hamon, 2020), la reconnaissance d'entités nommées (Dupont, 2017) et la classification de questions (Eshkol-Taravella *et al.*, 2022). Nous pouvons remarquer que les plongements lexicaux en tant que ressource sont relativement peu réutilisés. En effet, le fait qu'ils codent des informations linguistiques leur donne un statut de ressource lexicale mais leur utilisation s'avère différente des lexiques symboliques dans la mesure où ce sont davantage les modèles qui sont utilisés pour générer des plongements lexicaux propres à une application.

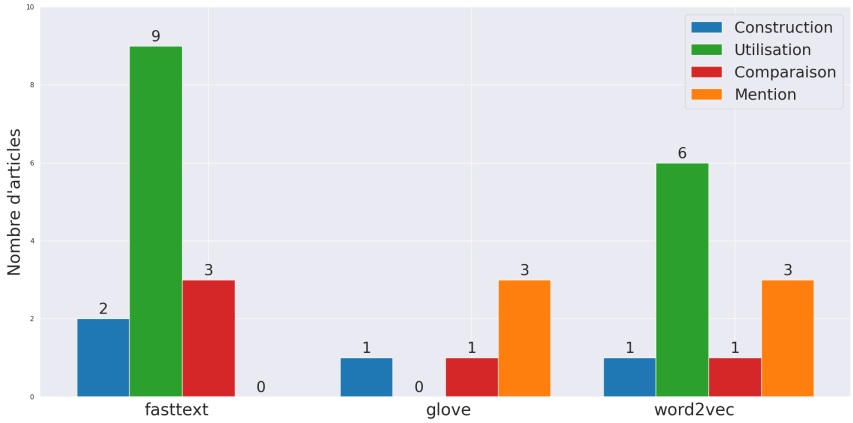

FIGURE 3 – Distribution des plongements lexicaux selon le type d'usage.

3.2 Réutilisation externe ou interne ?

Au-delà des applications, nous nous intéressons également aux personnes utilisant les ressources. Nous cherchons à déterminer si les utilisations d'une ressource sont menées par des personnes impliquées dans la construction de celle-ci. Pour ce faire, nous considérons une utilisation comme externe les cas où les auteurs de l'article ne font pas partie de la liste de créateurs donnée dans le tableau 1 [5].

Dans la figure 4, nous observons les proportions d'utilisations internes et externes pour les 15 ressources précédemment extraites. La figure montre que les ressources ont tendance à être réutilisées par des personnes extérieures à la ressource, à l'exception de `JeuxDeMots` où huit articles sur 13 comptent Mathieu Lafourcade dans leurs auteurs. Nous remarquons également que presque un tiers des utilisations du `Lefff` sont menées par un des créateurs de la ressource et que la moitié des utilisations du Lexique-Grammaire sont internes.

Observer si une ressource est utilisée par des personnes extérieures permet de questionner la réutilisabilité des ressources et ses facteurs déterminants. Inspirés par Cohen *et al.* (2005) décrivant les critères de réutilisabilité pour les corpus bio-médicaux, nous pouvons supposer que la taille de la ressource, sa couverture, son format, son âge ou la facilité de sa prise en main peuvent être des potentiels critères.

5. Nous précisons que les informations sur les concepteurs des ressources peuvent contenir des erreurs et que les articles RECITAL ne mentionnant pas les encadrants des étudiants peuvent avoir une influence sur les chiffres.

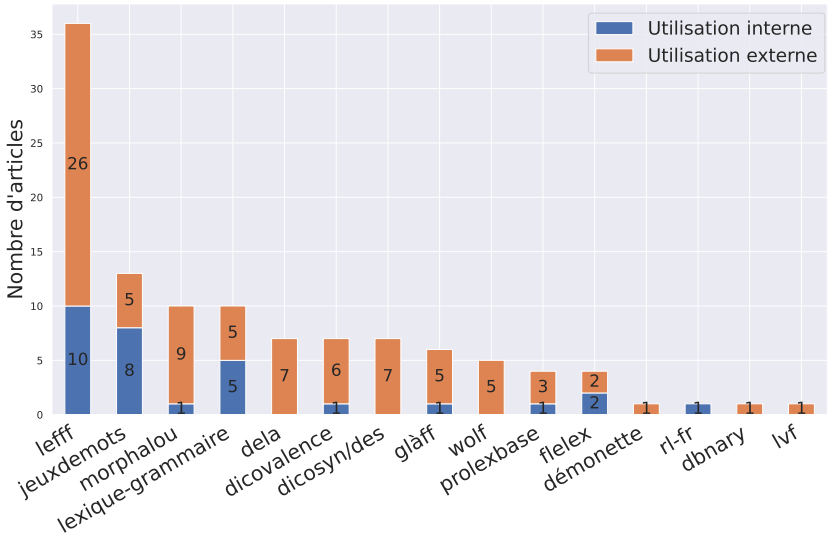

FIGURE 4 – Proportions de réutilisations internes ou externes pour les 15 ressources utilisées.

4 Discussion

À travers cette étude, nous observons que les ressources lexicales ont été utilisées de manière relativement stable depuis 20 ans. Un des exemples majeurs est le `Lefff`, utilisé près d'une fois par an depuis 2006 et considéré comme une ressource de référence pour le français. Elle doit ce statut notamment aux efforts de fusion de plusieurs ressources existantes (`Dicovalence`, `LVF`, `Lexique-Grammaire`) lui octroyant ainsi une meilleure couverture et complétude. Nous observons également que l'effet de l'âge a une influence sur nos chiffres, les ressources les plus récentes présentant une utilisation moindre. Par ailleurs, depuis plusieurs années, une nouvelle forme de ressource apparait avec les plongements lexicaux, de plus en plus présents du fait des récentes avancées des modèles de langue. Nous notons toutefois que les résultats ne montrent pas une massive utilisation des plongements lexicaux en tant que ressource, ces derniers étant davantage utilisés en tant qu'outil avec également la montée des modèles de langue tels que BERT (Devlin *et al.*, 2019). Bien que ces modèles de langue connaissent un fort succès depuis une dizaine d'années, des recherches visent de plus en plus à y intégrer des ressources symboliques comme des lexiques ou des bases de connaissances dans le but d'encoder des informations linguistiques externes et ainsi améliorer leur interprétabilité (Roy & Pan, 2020; Yang *et al.*, 2021).

Au-delà de leur place dans le domaine du TAL, nous pouvons souligner que certaines ressources lexicales ne sont pas seulement destinées au TAL mais également à la linguistique (`Lexique4linguists` (Schalchli, 2022)) ou à la didactique du FLE avec par exemple la création de `FLELex` (François *et al.*, 2014).

En observant la dynamique des ressources lexicales dans le temps dans les actes de TALN, nous avons cherché à mettre en évidence des critères permettant une meilleure réutilisabilité. Toutefois, limités probablement par un faible échantillon (moins de 300 articles) et des pertes durant la phase de conversion, nous ne pouvons noter aucune réelle tendance dans cette étude qui mériterait d'être étendu à une conférence plus conséquente telle que la conférence internationale de ressources langagières, LREC. Tous les scripts et annotations relatifs à l'article sont mis à disposition [6].

6. https://gitlab.inria.fr/papers2/taln2023

Références

ATILF (2016). Tlfphraseo. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

ATILF (2019). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BALVET A., BARQUE L., CONDETTE M.-H., HAAS P., HUYGHE R., MARIN R. & MERLO A. (2011). La ressource nomade. confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus [the nomage resource. compare theoretical expectations with observations of linguistic behavior of nominalizations in corpus]. *Traitement Automatique des Langues*, **52**(3), 129–152.

BARANES M. (2012). Vers la correction automatique de textes bruités : Architecture générale et détermination de la langue d'un mot inconnu (towards automatic spell-checking of noisy texts : General architecture and language identification for unknown words) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 3 : RECITAL*, p. 95–108, Grenoble, France : ATALA/AFCP.

BAWDEN R. (2017). Machine translation of speech-like texts : Strategies for the inclusion of context. In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, p. 1–14, Orléans, France : ATALA.

BLACHE P. & RAUZY S. (2008). Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 282–291, Avignon, France : ATALA.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BURLOT F. & YVON F. (2018). Évaluation morphologique pour la traduction automatique : adaptation au français (morphological evaluation for machine translation : Adaptation to French). In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, p. 61–74, Rennes, France : ATALA.

CANDITO M., AMSILI P., BARQUE L., BENAMARA F., DE CHALENDAR G., DJEMAA M., HAAS P., HUYGHE R., MATHIEU Y. Y., MULLER P., SAGOT B. & VIEU L. (2014). Developing a French FrameNet : Methodology and First results. In *LREC - The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Islande. HAL : hal-01022385.

CHOI H.-S. (2022). État de l'art : Liage de ressources lexicales du français (state of the art : Linking French lexical resources). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, p. 55–68, Avignon, France : ATALA.

CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal : European Language Resources Association (ELRA).

COHEN K. B., FOX L., OGREN P. V. & HUNTER L. (2005). Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases : Mining Biological Semantics*, p. 38–45, Detroit : Association for Computational Linguistics.

CONSTANT M. & SIGOGNE A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, p. 49–56, Portland, Oregon, USA : Association for Computational Linguistics.

CONSTANT M. & TOLONE E. (2010). A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In M. D. GIOIA, Éd., *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie*, volume 1 de *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, p. 79–93. Aracne. ISBN 978-88-548-3166-7.

DANLOS L., NAKAMURA T. & PRADET Q. (2014). Vers la création d'un verbenet du français. In *Atelier FondamenTAL, TALN 2014*.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, **46**(4), 721–736. DOI : 10.1007/s10579-012-9193-0.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français.* Larousse-Bordas, Paris, France.

DUCEL F., FORT K., LEJEUNE G. & LEPAGE Y. (2022). Langues par défaut ? analyse contrastive et diachronique des langues non citées dans les articles de TALN et d'ACL (contrastive and diachronic study of unmentioned (by default ?) languages in TALN and ACL we study the application of the #BenderRule in natural language processing articles, taking into account a contrastive and a diachronic dimensions, by examining the proceedings of two NLP conferences, TALN and ACL, over time). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 144–153, Avignon, France : ATALA.

DUPONT Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique (feature exploration for French named entity recognition with machine learning). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, p. 42–55, Orléans, France : ATALA.

ENGUEHARD C., KANÉ S., MANGEOT M., MODI I. & SANOGO M. L. (2012). Vers l'informatisation de quelques langues d'afrique de l'ouest (towards the computerization of some west-african languages) [in French]. In *JEP-TALN-RECITAL 2012, Workshop TALAf 2012 : Traitement Automatique des Langues Africaines (TALAf 2012 : African Language Processing)*, p. 27–40, Grenoble, France : ATALA/AFCP.

ESHKOL-TARAVELLA I., BARBEDETTE A., LIU X. & SOUMAH V.-G. (2022). Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral (automatic classification of spontaneous vs). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 305–314, Avignon, France : ATALA.

FAUCONNIER J.-P. (2015). French word embeddings.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

FRANCOIS T., BILLAMI M. B., GALA N. & BERNHARD D. (2016). Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension (automatic ranking of synonyms according to their reading and comprehension difficulty). In *Actes de la*

conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Articles longs), p. 15–28, Paris, France : AFCP - ATALA.

FRANÇOIS T., GALA N., WATRIN P. & FAIRON C. (2014). FLELex : a graded lexical resource for French foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande. HAL : hal-01758123.

GALA N. (2013). Ressources lexicales mono- et multilingues : une évolution historique au fil des pratiques et des usages. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 de *Lingvisticæ Investigationes Supplementa*, p. 1–42. John Benjamins Publishing. HAL : hal-03203895.

GALA N., FRANÇOIS T. & FAIRON C. (2013). Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLex - Electronic Lexicography*, Tallinn, Estonie. HAL : hal-03194427.

GOSSET C., BOUMEDYEN BILLAMI M., LAFOURCADE M., BORTOLASO C. & DERRAS M. (2021). Extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie dans un corpus métier (automatic extraction of hypernym and hyponym relations in a professional corpus). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 162–170, Lille, France : ATALA.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon : European Language Resources Association (ELRA).

GROSS M. (1975). *Méthodes en syntaxe*. Hermann, Paris, France.

HATHOUT N. & NAMER F. (2014). Démonette, a french derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**(5), 125–168.

KHALDI H., ABDAOUI A., BENAMARA F., SIGEL G. & AUSSENAC-GILLES N. (2020). Classification de relations pour l'intelligence économique et concurrentielle (relation classification for competitive and economic intelligence). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 27–39, Nancy, France : ATALA et AFCP.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LAFOURCADE M. & LE BRUN N. (2020). Jeuxdemots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. *Revue Lexique*, **27**, 47–86.

LAFOURCADE M., LE BRUN N. & ZAMPA V. (2014). Les couleurs des gens. In *TALN : Traitement Automatique des Langues Naturelles*, Marseille, France. HAL : lirmm-01471671.

LOPEZ P. (2008–2023). Grobid. https://github.com/kermitt2/grobid.

LUX-POGODALLA V. (2014). Integrating lexicographic examples in a lexical network (intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, p. 586–591, Marseille, France : Association pour le Traitement Automatique des Langues.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In *First InternationalWorkshop on Lexical Resources, WoLeR 2011*, p. 54–61, Ljubljana, Slovénie. HAL : hal-00686467.

MAKS I., IZQUIERDO R., FRONTINI F., AGERRI R., VOSSEN P. & ANDONI AZPEITIA (2014). Generating polarity lexicons with wordnet propagation in 5 languages. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande : European Language Resources Association (ELRA).

MANGEOT M. & CHALVIN A. (2006). Dictionary building with the jibiki platform : the GDEF case. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genève, Italie.

MANGEOT-NAGATA M. (2016). Collaborative Construction of a Good Quality, Broad Coverage and Copyright Free Japanese-French Dictionary. *International Journal of Lexicography*, **31**(1), 78–112. DOI : 10.1093/ijl/ecw035, HAL : hal-01712271.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

MARIANI J., FRANCOPOULO G. & PAROUBEK P. (2019a). The nlp4nlp corpus (i) : 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00036.

MARIANI J., FRANCOPOULO G., PAROUBEK P. & VERNIER F. (2019b). The nlp4nlp corpus (ii) : 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00037.

MERTENS P. (2010). Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE Conversion vers un format utilisable en TAL. In *Conference Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, p. 746–751.

MISSUD A., AMSILI P. & VILLOING F. (2020). VerNom : une base de paires morphologiques acquise sur très gros corpus (VerNom : a French derivational database acquired on a massive corpus). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 305–313, Nancy, France : ATALA et AFCP.

MORLANE-HONDÈRE F. & FABRE C. (2012). Étude des manifestations de la relation de méronymie dans une ressource distributionnelle (study of meronymy in a distribution-based lexical resource) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 169–182, Grenoble, France : ATALA/AFCP.

NASR A., BÉCHET F., REY J.-F., FAVRE B. & LE ROUX J. (2011). MACAON an NLP tool suite for processing word lattices. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, p. 86–91, Portland, Oregon : Association for Computational Linguistics.

PENNINGTON J., SOCHER R. & MANNING C. (2014). GloVe : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543. DOI : 10.3115/v1/D14-1162.

PLOUX S. & VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL*, **39**, 161–182. HAL : halshs-00009433.

PROST J.-P. (2022). Integrating a phrase structure corpus grammar and a lexical-semantic network : the HOLINET knowledge graph. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 613–622, Marseille, France : European Language Resources Association.

RANDRIATSITOHAINA T. & HAMON T. (2020). Identification des problèmes d'annotation pour l'extraction de relations (identification of annotation problem for the relation extraction). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 323–331, Nancy, France : ATALA et AFCP.

RAUZY S. & BLACHE P. (2007). Un lexique syntaxique des verbes du français : VfrLPL. 7 pages.

ROHATGI S. (2022). Acl anthology corpus with full text. Github.

ROY A. & PAN S. (2020). Incorporating extra knowledge to enhance word embedding. In C. BESSIERE, Éd., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, p. 4929–4935 : International Joint Conferences on Artificial Intelligence Organization. Survey track, DOI : 10.24963/ijcai.2020/686.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte. HAL : inria-00521242.

SAGOT B. (2019). Développement d'un lexique morphologique et syntaxique de l'ancien français (development of a morphological and syntactic lexicon of Old French). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, p. 265–274, Toulouse, France : ATALA.

SAGOT B. & FIŠER D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Maroc.

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2006). Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 288–297, Leuven, Belgique : ATALA.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SCHALCHLI G. (2022). Lexique4linguists. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

SÉRASSET G. (2015). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, **6**(4), 355–361. DOI : 10.3233/SW-140147, HAL : hal-00953638.

STRNADOVÁ J. & SAGOT B. (2011). Construction d'un lexique des adjectifs dénominaux (construction of a lexicon of denominal adjectives). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, p. 67–72, Montpellier, France : ATALA.

TRAN M. & MAUREL D. (2006). Prolexbase - un dictionnaire relationnel multilingue de noms propres. *Trait. Autom. des Langues*, **47**(3), 115–139.

TROUILLEUX F. (2012). Le DM, a French Dictionary for NooJ. In B. B. KRISTINA VUČKOVIĆ & M. SILBERZTEIN, Éds., *Automatic Processing of Various Levels of Linguistic Phenomena : Selected Papers from the NooJ 2011 International Conference*, p. 16–28. Cambridge Scholars Publishing. HAL : hal-00702348.

VAN DEN EYNDE K. & MERTENS P. (2006). Le dictionnaire de valence dicovalence : manuel d'utilisation.

VAN DEN EYNDE K. & MERTENS P. (2010). Le dictionnaire de valence dicovalence : manuel d'utilisation version 2.0.

VILLEGAS M., MELERO M., BEL N. & GRACIA J. (2016). Leveraging RDF graphs for crossing multiple bilingual dictionaries. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 868–876, Portorož, Slovénie : European Language Resources Association (ELRA).

YANG J., XIAO G., SHEN Y., JIANG W., HU X., ZHANG Y. & PENG J. (2021). A survey of knowledge enhanced pre-trained models. *arXiv preprint arxiv :2110.00269*.

Annexes

Type	Ressources LRE map	Ressources Ortolang
Lexiques monolingues	Dicovalence, Lefff, JeuxDeMots, LGLex, French FrameNet, LVF, Lexique-Grammaire, VerbeNet, ReSyf, WOLF, GLAFF, DELA, FLELex, Démonette	Démonette, Morphalou, DES, Lexique4linguists, Holinet, DM, RL-Fr, VerNom, Nomage, Prolexbase, Dicovalence, TLFPhraseo, MarsaLex, VfrLPL
Lexiques multilingues	Apertium RDF Graph, OpeNER-sentiment-lexicons, DBnary, DiLAF	Prolexbase
Lexiques non sélectionnés	Wikitionary, FreeLang, Free dictionary download, Terminesp LD	Dictionnaire informatisé des Mots d'Affect, DSR, Termes de base du diagnostic orthophonique, LGeRM
Lexiques non disponibles (liens cassés)	JournalisticNL11, EUROSENTIMENT, CorpusDRF, DeQue, tl_dv2_ladl par-lvf, V2R, MotaMot, Multilingual glossary of technical and popular medical terms, Bilingual Dictionaries, Verbnet like classification of French verbs	
Doublons	Lefff (x2), LGLex 3.3 (x2), Dicovalence 2, LG, FLELLex, WOLF, DBnary (x2)	
Total	44	19

TABLE 2 – Lexiques du français libres d'utilisation présents dans la LRE map et Ortolang.

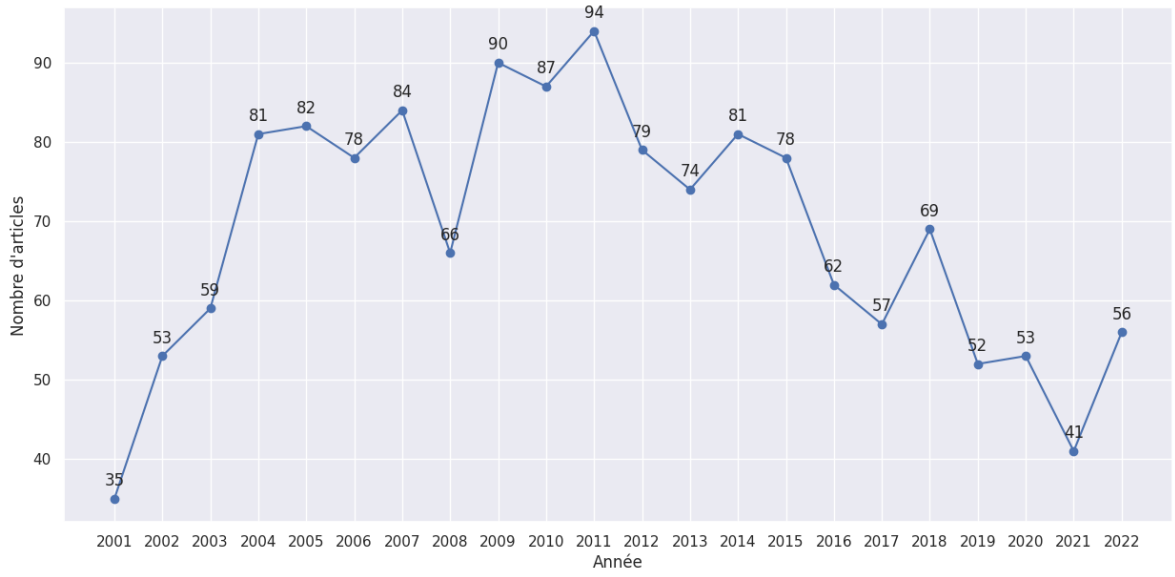

FIGURE 5 – Nombres d'article du corpus TALN 2001-2022 par année.