CONSTITUTION DE CORPUS EN MORPHOLOGIE : L'UTILISATION DU WEB COMME RESSOURCE

L'essor du numérique et du Web ont provoqué « une révolution copernicienne » (Bergounioux & Dal, 2016 : 7) en morphologie. Les ressources lexicographiques ne sont plus le seul objet d'étude du morphologue. Cette révolution donne lieu à des questionnements, comme en témoignent les travaux de Lüdeling *et al.* (2007), Fradin *et al.* (2008), Dal & Namer (2015), pour ne citer qu'eux. L'essence même de la morphologie se voit interrogée tant le champ des possibles en matière de corpus s'étend avec l'avènement du Web. Quel est le but du morphologue ? Quelles sont les avantages et inconvénients d'un tel changement ? Quelles données le morphologue trouve-t-il sur le Web ? Sont-elles toutes valides ? Dans cette présentation, nous proposerons de répondre à ces questions en présentant le protocole méthodologique mis en œuvre pour construire notre corpus.

Notre objectif est d'analyser les mots morphologiquement construits sur 90 Noms propres de Personnalités Politiques françaises (désormais NPP), sélectionnées pour avoir occupé une fonction de premier plan depuis 1980 (*ex.* ALAIN JUPPÉ$_{NPr}$ > JUPPÉISTE$_{Nc}$).

Dans un premier temps, nous montrerons que trois singularités propres à notre sujet nous ont orientés vers le Web. **1)** Le caractère contemporain des politiques sélectionnés fait que les mots qui en découlent ne sont pas enregistrés dans les dictionnaires et les exclut des corpus constitués il y a plusieurs années (*ex.* FrWaC). **2)** Certains mots construits sur NPP ont un caractère éphémère, voire tellement original (*ex.* FRANÇOIS BAYROU$_{NPr}$ > BAYROUBUS$_{Nc}$), qu'ils ne figureront jamais dans une nomenclature dictionnairique (sur la notion d'*occasionnalisme* : Hohenhaus, 2005 ; Dal & Namer, 2016). **3)** Travailler sur une grande quantité de données, accessibles grâce au Web, *i.e.* dans une démarche extensive (*cf.* Hathout *et al.*, 2008), nous permet de dresser une cartographie des différents types de construction morphologique opérant sur NPP et d'analyser suffisamment d'occurrences pour dégager des nouveautés (*ex.* Hathout *et al.*, 2009).

Nous présenterons ensuite les inconvénients rencontrés lors l'utilisation du Web. **a)** Les moteurs de recherche ne permettent plus d'effectuer des requêtes complexes et automatisées, *i.e.* d'utiliser des opérateurs booléens ou de chercher une liste de mots (*ex.* outil WaliM, Namer, 2013). **b)** Une collecte à partir du Web nécessite un post-traitement pour constituer un *corpus* (au sens de Sinclair, 1996), *i.e.* éliminer le bruit et annoter les données. **c)** L'hétérogénéité des sources soulève également la question de la validité des données du Web.

Enfin, nous montrerons que nous avons pu élaborer une méthodologie de constitution de corpus adaptée, en collaboration avec l'entreprise Data-Observer[1], afin de pallier ces problèmes. Somme toute, nous produisons une analyse empirique, basée sur des données authentiques, *i.e.*

[1] Data-Observer (www.data-observer.com) est une startup spécialisée dans la collecte, le traitement et l'analyse des données textuelles issues du Web.

effectivement produites par des locuteurs, répondant, selon nous, à l'objectif du morphologue : décrire l'usage (*cf.* Dal & Namer, 2012).

BIBLIOGRAPHIE

Baroni M., Bernardini S., Ferraresi A. & 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.

Bergounioux G. & Dal G. (2016). Les observables entre théorie et technologie. Deux exemples : la création lexicale et les amorces. *Le Français Moderne – Revue de linguistique française, CILF* (Conseil international de la langue française), 13-36.

Dal G. & Namer F (2016). À propos des occasionnalismes. *SHS Web of Conference 27*, CMLF 5, 1-18.

Dal G. & Namer F. (2012). Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie. *SHS Web of Conferences 1*, CMLF 3, 1261-1276.

Dal G. & Namer F. (2015). 133. Internet. In: Peter O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (ed.), *Word Formation – An International Handbook of the Languages of Europe*, Volume 3. Berlin/New York: De Gruyter Mouton, 2372-2386.

Fradin B., Dal G., Grabar N., Namer F., Lignon S., Tribout D. & Zweigenbaum P. (2008). Remarques sur l'usage des corpus en morphologie. *Langages* 171 (3), 34-59.

Hathout N., Montermini F. & Tanguy L. (2008). Extensive data for morphology: using the World Wide Web. *Journal of French Language Studies* 18, 67-85.

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

Hohenhaus P. (2005). Lexicalization and Institutionalization. In: P. Štekauer & R. Lieber (eds), *Handbook of Word-Formation*. Berlin, Dordrecht, Heidelberg, Norwell: Springer, 353-373.

Lüdeling A., Stefan E. & Baroni M. (2007). Using Web data for linguistic purposes. In: M. Hundt, N. Nesselhauf & C. Biewer (eds), *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi, 7-24.

Namer F. (2013). WaliM : valider les unités morphologiquement complexes par le Web. In : G. Dal & D. Amiot, *Repères en morphologie*, 171-181 [http://stl.recherche.univ-lille3.fr/textesenligne/Reperes-Morphologie/Namer_Reperes_morphologie_p171-181.pdf].
Réédition en ligne du laboratoire STL de Namer F. (2003). In : B. Fradin, G. Dal, N. Hathout, F. Kerleroux, M. Plénat & M. Roché (éd.), *Silexicales 3 : Les unités morphologique*s. Villeneuve d'Ascq : Presses Universitaires du Septentrion, 142-150.

Sinclair J. (1996). *Preliminary recommendations on Corpus Typology*. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standars). En ligne : http://www.ilc.cnr.it/EAGLES96/corpustyp/node5.html#SECTION00041000000000000000.

RESSOURCE CITÉE

- FrWaC (*cf.* Baroni *et al.*, 2009), accessible en ligne : http://nl.ijs.si/noske/wacs.cgi/first_form?corpname=frwac;lemma=grave;lpos=