Préface

Des modèles linguistiques aux réalisations informatiques : quand les frontières bougent en morphologie

En linguistique, et à plus forte raison en morphologie, ces quarante dernières années ont vu revenir au devant de la scène des principes théoriques et descriptifs qu'avaient partiellement occultés plusieurs décennies d'hégémonie du courant générativiste. Plusieurs événements ont directement ou indirectement, contribué à ce revirement, la période charnière étant la reconnaissance, de la part de Noam Chomsky, de l'indépendance de la morphologie, qu'il n'assimile plus, à partir de (Chomsky 1970), à une branche de la syntaxe :

– la remise à l'honneur d'approches fondées sur l'analogie proportionnelle (pour un historique de la notion, voir (Lepage, 2003 : 37-78), partiellement repris dans (Dal 2003, 2008) où l'accent est mis sur le rôle de l'approche en morphologie) ;

– l'intervention de facteurs psycholinguistiques, qui privilégient le point de vue du locuteur dans la création lexicale (Bybee, 1985 ; Blevins, 2006 ; Burzio, 2002) ;

– l'évolution des capacités de stockage des ordinateurs et l'extension de l'Internet qui rendent possibles la découverte et l'étude des données réelles et massives provenant de corpus écrits ou retranscrits de l'oral, dans un nombre important de langues ou dans une grande variété de registres langagiers.

Au final, les travaux en morphologie de nos jours font intervenir des raisonnements qui se retrouvent de part et d'autre d'une ligne de partage entre d'un côté un courant morphologique compositionnel, morphématique, inspiré du cadre génératif, et de l'autre un courant réalisationnel, inférentiel, qui considère les relations morphologiques sous un angle paradigmatique, et qui cherche à rendre compte des données dans leur diversité (Stump, 2001). Cette ligne de partage ne comporte pas de frontière nette mais offre des situations intermédiaires. Aux courants théoriques qui coexistent de nos jours correspondent des conceptions différentes pour :

– les unités en jeu : morphèmes, morphèmes et lexèmes, lexèmes, mots ;

– les mécanismes de construction : règles binaires orientées, prescriptives, et assorties d'exceptions, ou patrons descriptifs non orientés et paradigmatiques ;

– ce qui est mis en exergue dans une construction : l'entrée de la règle (et le pouvoir génératif de celle-ci) ou sa sortie, et les différents modèles auxquels a réellement accès le locuteur.

Une ligne de partage similaire existe en traitement automatique des langues. D'une part, depuis le début des années 1990, les méthodes statistiques occupent le haut du pavé en TAL (Charniak, 1993 ; Church et Mercier 1993 ; Manning et Schütze, 1999 ; Church, 2011). Ces méthodes, fondées sur l'apprentissage automatique supervisé ou non supervisé ont connu de nombreux succès en étiquetage morphosyntaxique, en analyse syntaxique automatique et dans des tâches comme la traduction automatique (Brown *et al.*, 1993) ou la reconnaissance de la parole (Rabiner et Juang, 1993). D'autre part, de très nombreux systèmes de TAL restent fondés sur des descriptions linguistiques. En analyse syntaxique, par exemple, on observe que les résultats de ces systèmes sont supérieurs à ceux des analyseurs statistiques, ces derniers étant tributaires de la taille des corpus arborés disponibles. Dans le domaine de la morphologie informatique, la situation est globalement la même avec cependant une différence notable. Il n'existe en effet en morphologie qu'une seule ressource comparable à un corpus arboré. Il s'agit de la base de données CELEX (Baayen *et al.*, 1995) qui comporte une description d'un fragment significatif du lexique de l'anglais, de l'allemand et du néerlandais. Cette ressource n'a cependant pas été utilisée de manière intensive pour l'entraînement de systèmes d'analyse morphologique statistique. La raison en est qu'en morphologie computationnelle, les systèmes statistiques qui sont développés utilisent des techniques d'apprentissage non supervisées (Hammarström et Borin, 2011), alors qu'en syntaxe computationnelle, les techniques d'apprentissage mises en œuvre sont supervisées. Le nombre important de systèmes de traitement de la morphologie fondés sur l'apprentissage non supervisé est également lié aux campagnes d'évaluation Morpho-Challenge. Les systèmes morphologiques qui participent à ces compétitions doivent réaliser une analyse morphématique des mots d'un corpus de textes bruts. Ces campagnes ont connu un succès réel puisque que plus de cinquante systèmes y ont participé pour la période 2005-2010. Elles ont de ce fait une visibilité très importante et occupent une place centrale dans le champ. L'un des articles présentés dans ce numéro y est consacré.

Cependant, il existe aussi en morphologie computationnelle un ensemble significatif de travaux dont la visée est la modélisation de la morphologie et son étude au travers de modèles à large couverture. Ces modèles sont des systèmes symboliques qui implémentent des descriptions linguistiques réalisées dans le cadre de théories morphologiques ; c'est notamment le cas pour l'analyseur DériF (Namer, 2009), qui implémente un formalisme inspiré du courant lexématique de la morphologie lexicale (Aronoff, 1976, 1994 ; Corbin 1987 ; Fradin, 2003).

Le présent numéro de la revue TAL présente un ensemble de travaux qui se situent à cheval sur plusieurs des zones dessinées par ces lignes de partage et qui

illustrent chacun une facette de la morphologie informatique et ses évolutions possibles.

L'article de J.-F. **Lavallée et P. Langlais** présente un travail réalisé dans le cadre d'une participation à Morpho-Challenge. L'originalité du système présenté, *Moranapho*, repose sur le fait qu'il n'utilise pas une méthode statistique, mais qu'il est fondé sur la mise en évidence de régularités générales dans la morphologie des langues. Cet analyseur, indépendant des langues, parvient à produire une analyse morphématique en utilisant l'analogie formelle. C'est donc à la fois un système symbolique (ce qui l'oppose aux méthodes statistiques), une approche non fondée sur l'emploi de règles (ce qui le distingue des raisonnements d'inspiration générativiste), et un analyseur conduisant à une décomposition en morphèmes (ce qui en fait un candidat performant dans les campagnes de Morpho-Challenge). Parmi les résultats qualitatifs remarquables que l'article donne à voir, et qui montrent leur supériorité sur des approches concaténatoires, nous pouvons citer la prise en compte des phénomènes d'apophonie, comme l'alternance vocalique (Umlaut) caractéristique des langues germaniques (*lang/länger*).

L'article de **S. Virpioja, V. T. Turunen, S. Spiegler, O. Kohonen et M. Kurimo** est lui aussi consacré à Morpho-Challenge. Cependant, il ne décrit pas une participation particulière à l'une des éditions de la compétition, mais propose une comparaison de différentes mesures d'évaluation des analyseurs morphologiques. En d'autres termes, il s'agit d'une méta-évaluation dans laquelle les auteurs utilisent les résultats des systèmes qui ont participé aux éditions de Morpho-Challenge pour comparer trois mesures principales et cinq de leurs variantes. La plus ancienne de ces trois mesures, nommée MC, a été utilisée à partir de l'édition 2007 de la compétition. Cette mesure est fondée sur une comparaison des arcs d'un graphe de cooccurrences dans lequel deux mots sont connectés s'ils partagent un radical ou un affixe. MC présente plusieurs propriétés intéressantes, dont une bonne corrélation aux résultats d'évaluation dans des tâches de TAL, comme la recherche d'information ou la traduction automatique. Un autre point fort de MC réside dans sa faible complexité computationnelle, ce qui lui permet de traiter des corpus d'évaluation relativement volumineux. Cette mesure souffre en revanche d'un manque de fiabilité : les systèmes évalués peuvent améliorer artificiellement leur rappel en ajoutant à leurs résultats des découpages supplémentaires. Pour pallier cet inconvénient, Spiegler et Monson (2010) ont proposé EMMA, une seconde métrique qui consiste à apparier un à un chacun des morphèmes qui composent les découpages candidat et de référence. Cette mesure résout efficacement le problème de fiabilité tout en gardant une bonne corrélation avec les performances dans les tâches de TAL. Mais EMMA présente à son tour un point faible : sa complexité computationnelle ne lui permet pas de traiter des échantillons de plus de 2 000 mots. Les auteurs de l'article proposent donc une troisième métrique, adaptée de EMMA, où la correspondance 'un-à-un' de cette dernière est remplacée par deux correspondances 'plusieurs-à-un'. Les auteurs montrent que cette nouvelle mesure

satisfait les trois critères de qualité identifiés : elle a une bonne corrélation avec les évaluations par les tâches ; elle est fiable ; elle peut être utilisée avec des corpus de taille importante. Ce compromis est cependant obtenu au prix d'une dégradation dans l'interprétabilité des correspondances, la correspondance 'un-à-un' permettant une meilleure compréhension des résultats des systèmes évalués.

En dehors de ces deux articles qui sont directement connectés à Morpho-Challenge, le numéro propose trois autres contributions qui montrent la variété des approches et des intérêts en morphologie informatique. Là encore, on observe que les positionnements sont complexes et ne permettent pas de superposer les orientations des auteurs avec le partitionnement induit par les lignes de partage invoquées ci-dessus, qui opposent d'une part traitement symbolique et traitement statistique, et de l'autre règles d'agencement de morphèmes et analyses fondées sur les contraintes. Les trois articles illustrent en effet trois voies qui indiquent ce que pourraient être les futures orientations de la recherche en morphologie informatique.

Le premier de ces articles est consacré à la modélisation de nouvelles avancées théoriques en morphologie flexionnelle. Dans le cadre de recherches en typologie des langues, qui connaissent un essor important depuis quelques années (Corbett, 2007, 2010), ces avancées mettent l'accent sur l'écart qui existe entre idéal et réalité. Une règle morphologique idéalisée ou canonique est à la fois transparente formellement, parfaitement reproductible à l'identique, et sémantiquement compositionnelle. Alors que cette vision de la morphologie est celle qu'enseignent les manuels, et que modélise le courant morphématique, elle est mise à mal par la réalité des langues du monde, dont l'étude montre que la canonicité n'est jamais atteinte. C'est sur ces écarts au canon que porte l'article de **G. Walther et B. Sagot**. En s'appuyant sur des phénomènes variés appartenant à des langues aussi diverses que le français, le latin, l'italien, le kurde, le persan, le croate et le slovaque, et ayant comme point commun celui d'enfreindre les critères de canonicité, cet article présente *PARSLI*, un modèle d'analyse flexionnelle, ainsi que son implémentation. Cette modélisation est une parfaite illustration des travaux actuels en TAL, où les outils et les expérimentations ne portent plus sur des ensembles prototypiques. Le travail de ces auteurs a été en effet réalisé sur un lexique de grande taille, et permet d'avoir une idée claire de l'adéquation du modèle aux données réelles.

L'article de **C. Celata, B. Calderone et F. Montermini** illustre une autre des thématiques qui pourraient connaître, dans un avenir proche, un développement notable : la modélisation computationnelle du traitement cognitif de la morphologie des langues. La modélisation des processus cognitifs a toujours fait partie des thématiques du TAL. Elle a connu une grande activité dans les années 1980 et 1990 dans le sillage de l'intelligence artificielle (IA). Avec le développement des méthodes statistiques, l'IA, et avec elle la modélisation des processus cognitifs, ont beaucoup perdu de leur popularité. Dans le domaine de la morphologie, les travaux de modélisation ont essentiellement été réalisés par les psycholinguistes, dans le but notamment de départager différents modèles : décompositionnels, d'appréhension

globale, ou mixtes. Les plus connus sont certainement les travaux de Rumelhart et McClelland (1986). C'est dans leur lignée que s'inscrit cet article. Il s'agit d'une recherche pluridisciplinaire qui combine deux dispositifs expérimentaux. Le premier est un ensemble d'expériences psycholinguistiques destinées à vérifier que le traitement morphologique des mots met en jeu des unités infralexicales et que la position de ces unités dans le mot a une incidence forte sur leur identification. Elle confirme que la position finale joue un rôle plus important que la position initiale dans cette identification. Ces expériences sont réalisées sur l'italien en utilisant des pseudo-mots afin d'annuler tout effet sémantique. Parallèlement, un modèle computationnel de la phonotactique de l'italien a été réalisé et ses résultats ont été comparés à ceux des sujets qui ont participé aux expériences psycholinguistiques. Ce modèle représente le lexique au moyen d'une carte auto-organisatrice de Kohonen (1995) qui rapproche les mots qui présentent des propriétés phonotactiques similaires. Tout comme les locuteurs de l'italien, ce modèle est capable d'apprendre que les phonèmes qui se trouvent en position finale sont plus discriminants que ceux qui se trouvent en position initiale. Le modèle a été utilisé pour calculer la similarité entre les pseudo-mots qui ont été soumis aux sujets dans l'expérience psycholinguistique. Les résultats du modèle présentent une corrélation forte avec les jugements de similarité des sujets confirmant ainsi la plausibilité psychologique de ce dernier.

Enfin, l'article de **J. Chmielik et N. Grabar** présente une étude dont le but est d'estimer la contribution des informations morphologiques dans la caractérisation de textes dans le domaine médical. Là encore, le travail est expérimental, et, là encore, le modèle morphologique utilisé rompt avec les approches privilégiées dans Morpho-Challenge. En se servant des résultats de l'analyseur DériF (Namer, 2009), les auteurs montrent que les indices morphologiques constituent des marqueurs supérieurs (en précision et en rappel) aux indices lexicaux pour classer les textes médicaux suivant leur degré de spécialisation. L'article témoigne ainsi du rôle que peut jouer l'analyse morphologique en typologie des textes. Le système d'analyse morphologique choisi privilégie une vision linguistiquement contrainte et non concaténative de la morphologie computationnelle, ce qui se traduit par une portée des affixes et des composants qui varie en fonction de la sémantique des mots analysés. Le caractère original du travail présenté porte donc non seulement sur la finalité de l'expérience menée, mais concerne également la méthode adoptée pour réaliser cet objectif.

En somme, l'objectif initial de ce numéro était de présenter la morphologie automatique sous l'angle de ses frontières et interactions. Les articles présentés remplissent avantageusement, nous semble-t-il, cette tâche à facettes multiples. Les uns ou les autres montrent en effet que les systèmes d'analyse morphologique constituent des variantes sur une échelle dont les pôles sont l'analyse non supervisée, purement séquentielle d'une part, et les modèles computationnels linguistiquement motivés, de l'autre. Les articles donnent ainsi à voir un panorama

de théories linguistiques implémentées, allant de l'analogie à la règle, en passant par les systèmes de contraintes. Ils franchissent les frontières en matière de langues traitées, et d'objectifs pratiques visés. Ils illustrent enfin l'interaction entre la morphologie, la grammaire, la classification documentaire et la psycholinguistique.

À l'heure actuelle, les travaux et les réalisations les plus centrales en morphologie computationnelle se consacrent aux approches fondées sur l'apprentissage non supervisé et aux campagnes d'évaluation Morpho-Challenge. Pour faire écho au succès de ces dernières, nous proposons au lecteur, avec l'article de Virpioja *et al.*, un bilan permettant de mieux connaître les mesures qui sont utilisées pour évaluer les analyseurs morphologiques et qui deviendront, à terme, des normes *de facto* similaires à BLEU (Papineni, 2002) pour la traduction automatique. Mais notre objectif est aussi d'éclairer les zones plus périphériques en traitement automatique de la morphologie. Les articles de ce volume montrent que la morphologie computationnelle s'intéresse à des espaces autres que la hiérarchisation des systèmes de décomposition en segments en fonction de leur efficacité, et que différentes disciplines et applications sont demandeuses d'approches nouvelles.

Ainsi, nous avons souhaité orienter ce numéro pour qu'il présente quelques-uns des apports mutuels entre les systèmes de traitement automatique en morphologie et :

– l'adoption de modèles linguistiques récents (les contraintes paradigmatiques ou l'analogie) ;

– l'application à des langues comportant une morphologie rétive à la décomposition par morphèmes car mettant en jeu une flexion non canonique ;

– l'interaction avec des disciplines connexes (la psycholinguistique) ;

– l'ouverture à de nouvelles applications (le classement documentaire).

L'intersection entre modèles linguistiques, approches cognitives et traitement automatique est encore peu connue. Nous espérons, par la variété des articles qui constituent ce numéro, avoir montré la richesse potentielle des résultats auxquels on peut s'attendre en morphologie computationnelle, résultats qui mettent en jeu des rapports de sens, confrontent des hypothèses fondées sur la complexité des structures et projettent les besoins et les capacités des locuteurs.

Références

Aronoff M., *Word Formation in Generative Grammar*, Cambridge, The MIT Press, 1976.

Aronoff M., *Morphology by Itself*, Cambridge, MIT Press, 1994.

Baayen R. H., Piepenbrock R., Gulikers L., *The CELEX Lexical Database (CD-ROM)*, Philadelphia, PA, Linguistic Data Consortium, University of Pennsylvania, 1995.

Blevins J. P., « Word-based morphology », *Linguistics* 42, p. 531-573, 2006.

Brown P. F., Della Pietra V. J., Della Pietra S. A., Mercer R. L., « The Mathematics of Statistical Machine Translation: Parameter Estimation », *Computational Linguistics*, vol. 19, n° 2, p. 263-311, 1993.

Burzio L., « Surface-to-surface morphology: when your representations turn into constraints ». In *Many Morphologies*, P. Boucher (ed.), Somerville, Cascadilla Press, p. 142-177, 2002.

Bybee J., *Morphology: A study of the relation between meaning and form*, Philadelphia, Benjamins, 1985.

Charniak E., *Statistical Language Learning*, Cambridge Mass, MIT Press, 1993.

Chomsky N. « Remarks on Nominalization », In *Readings in English Transformational Grammar*, R. A. Jacobs and P. S. Rosenbaum (eds.), Waltham Mass, Ginn, p. 184-221, 1970.

Corbett G., « Canonical typology, suppletion and possible words », *Language* 83, p. 8-42, 2007.

Corbett G., « Canonical Derivational Morphology », *Word Structure* 3, p. 141-155, 2010.

Corbin D., *Morphologie dérivationnelle et structuration du lexique.* Lille, Presses Universitaires de Lille, 1987.

Church K., « A pendulum swung too far » *Linguistic Issues in Language Technology*, vol. 6, n° 5, 2011.

Church K., Mercer R., « Introduction to the Special Issue on Computational Linguistics Using Large Corpora », *Computational Linguistics*, vol. 19, n° 1, p. 1-24, 1993.

Dal G., « Analogie et lexique construit : quelles preuves ? », *Cahiers de grammaire* 28, p. 9-30, 2003.

Dal G., « L'analogie dans le domaine du lexique construit : un retour ? » *1er Congrès mondial de linguistique française*, 2008, Paris, ILF, p. 1575-1587.

Fradin B., *Nouvelles approches en morphologie*. Paris, Presses Universitaires de France, 2003.

Hammarström H., Borin L., « Unsupervised learning of morphology », *Computational Linguistics*, vol. 32, n° 2, p. 309-350, 2011.

Kohonen T., *Self-Organizing Maps*, Springer Verlag, Berlin/Heidelberg, 1995.

Lepage Y., De l'analogie rendant compte de la communication en linguistique, Mémoire d'habilitation à diriger des recherches, Université de Grenoble1, 2003.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, Cambridge, MIT Press, 1999.

Namer F., *Morphologie, Lexique et TAL : l'analyseur DériF*, London, Hermes Sciences Publishing, 2009.

Papineni K., Roukos S., Ward T., Zhu W. J., « BLEU: a method for automatic evaluation of machine translation », *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 311-318, 2002.

Rabiner L., Juang B.-H., *Fundamentals of speech recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

Rumelhart D. E., MacClelland J. L., « On learning the past tenses of English verbs. Implicit rules or parallele distributed processing? », in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vol. 2, J. L. MacClelland, D. E. Rumelhart (eds.), MIT Press, p. 216-261, 1986.

Spiegler S., Monson C., « EMMA: A Novel Evaluation Metric for Morphological Analysis », *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, Chine, 2010.

Stump G., *Inflectional Morphology: A Theory of Paradigm Structure,* Cambridge Studies in Linguistics 93, Cambridge, Cambridge University Press, 2001.

Remerciements

Nous tenons à exprimer notre gratitude envers le comité de lecture de la revue TAL. Nos plus vifs et chaleureux remerciements vont aux membres du comité spécifique formé pour ce numéro :

– Delphine Bernhard (LiLPa, Université Strasbourg, France)

– Olivier Bonami (LLF, Université Paris 4, France)

– Narjes Boufaden (KeaText & École de Technologie Supérieure de Montréal, Canada)

– Gilles Boyé (CLLE-ERSS, Université Bordeaux 3, France)

– Basilio Calderone (CLLE-ERSS, Université de Toulouse 2 Le Mirail, France)

– Bruno Cartoni (Département de Linguistique de Genève, Suisse)

– Walter Daelemans (CLiPS, Université d'Anvers, Belgique)

– Georgette Dal (STL, Université Lille 3, France)

– Hervé Déjean (Xerox Research Centre Europe, France)

– Joseph Dichy (ICAR, Université Lyon 2, France)

– Cécile Fabre (CLLE-ERSS, Université Toulouse 2, France)

– Dafydd Gibbon (Université de Bielefeld, Allemagne)

– Olivier Kraif (LIDILEM, Université Grenoble 3, France)

– Harald Hammarström (Radboud Universiteit Nijmegen, Pays-Bas et Max Planck Institute for Evolutionary Anthropology, Leipzig, Allemagne)

– Ania Kupsc (CLLE-ERSS, Université Bordeaux 3, France)

– Mikko Kurimo (Aalto University School of Science, Finlande)

– Jean-François Lavallée (CRIM, Université de Montréal, Canada)

– Stéphanie Lignon (ATILF, Université Nancy 2, France)
– Christian Monson (Oregon Health and Science University, USA)
– Vincent Ng (University of Texas at Dallas, USA)
– Lionel Nicolas (Molino de Ideas S.A., Espagne)
– Kemal Oflazer (Carnegie Mellon University, Qatar)
– Vito Pirrelli (ILC-CNR, Università degli Studi di Pavia, Italie)
– Nicolas Stroppa (Google Zürich, Suisse)
– Ludovic Tanguy (CLLE-ERSS, Université Toulouse 2, France)
– Evelyne Tzoukermann (ITS, Washington University, USA)
– Patrick Watrin (CENTAL, Université Catholique de Louvain, Belgique).

Nabil Hathout
CLLE-ERSS (UMR 5263) CNRS & Université de Toulouse-Le Mirail
Maison de la Recherche. F-31058 Toulouse cedex 9
nabil.hathout@univ-tlse2.fr

Fiammetta Namer
Nancy-Université & ATILF CNRS (UMR 7118)
CLSH – 23 bd Albert 1er – BP3397 – F-54015 Nancy Cedex
fiammetta.namer@univ-nancy2.fr