JOURNÉE D'ÉTUDE organisée dans le cadre de
l'ANR DEMONEXT - ANR-17-CE23-0005

SÉMANTIQUE POUR LES RESSOURCES EN MORPHOLOGIE DÉRIVATIONNELLE

Résumés

De la variété des paradigmes dérivationnels

Bernard Fradin, UMR LLF

Le fil directeur de cette présentation est de chercher à établir les critères que les paradigmes dérivationnels devraient remplir pour constituer des dispositifs permettant de faire des prédictions, à l'instar de ce qu'autorisent les paradigmes flexionnels. En s'aidant des travaux existant sur les paradigmes dérivationnels (Bauer 1997; Štekauer 2014), l'étude s'efforce de tirer au clair les diverses formes que revêt l'organisation conceptuelle de ceux-ci. Elle montre qu'il est fondé de distinguer plusieurs types de paradigmes dérivationnels, lesquels s'articulent suivant deux grandes distinctions : (i) être ancré dans un événement ou pas ; (ii) être centré sur l'humain ou pas. Les relations de contenu observées dans les paradigmes flexionnels semblent avoir trois sources : des schèmes enracinés dans des événements mettant généralement en jeu un agent ex. les réseaux d'action ou d'activité de (Roché 2017a); Roché (2017b); des propriétés associées à une entité que dénote le nom pivot d'un paradigme ex. 'pays, habitant, langue' ; enfin des méta-patrons de dérivation très fréquents et saillants au niveau de la langue même ex. la dérivation des verbes perfectifs et imperfectifs dans les langues slaves.

L'étude tire parti des travaux récents sur les paradigmes (Bonami & Beniamine 2016; Bonami & Strnadová 2018) pour fixer les critères qui peuvent faire d'un paradigme dérivationnel un outil de prédiction. Elle discute le cas des dérivés construits sur des noms d'animaux pour montrer comment s'organise un paradigme dont les cellules (nœuds) ne sont ni liés à un événement, ni centrés sur l'humain. Le volet empirique sera complété par la discussion des notions de réseau d'action et de réseau d'activité avancées par Roché (2017a, b) comme étant deux types de paradigmes dérivationnels majeurs. L'étude s'efforce d'évaluer l'intérêt de cette distinction.

Les paradigmes dérivationnels étant par nature fondés sur les relations sémantiques existant entre les éléments qu'ils relient, la solidité de cette fondation n'est pas absolue mais relative et doit conséquemment être étayée par des données statistiques. L'étude se bornera à indiquer les points pour lesquels le recours à ces données s'avère cruciale pour l'élaboration de paradigmes dérivationnels ayant une capacité prédictive.

Références

Bauer, Laurie. 1997. Derivational Paradigms. Yearbook of Morphology.243-56.

Bonami, Olivier & Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigm. Word Structure 9.156-82.

Bonami, Olivier & Jana Strnadová. 2018. Paradigm structure and predicability in derivational morphology.*Morphology*.

Roché, Michel. 2017a. Les familles dérivationnelles: comment ça marche? Université Toulouse 2 Jean Jaurès. ms.

Roché, Michel. 2017b. Un exemple de réseau constructionnel: ethnique, toponymes, gentilés. Toulouse: Université Toulouse 2 Jean Jaurès. ms.

Štekauer, Pavol. 2014. Derivational paradigms. The Oxford Handbook of Derivational Morphology, ed. by R. Lieber & P. Štekauer, 354-69. Oxford: Oxford University Press.

Les verbes dénominaux en français : existe-t-il une corrélation entre la massivité/comptabilité des noms bases et l'(a)télicité des verbes dérivés ?

Pauline Haas, Lattice UMR 8094 (CNRS, ENS/PSL, P3/USPC) & Pléiade EA 7338 (P13) & Université Paris 13, Rafael Marín STL UMR 8163 & Edwige Dugas STL UMR 8163 (CNRS, univ Lille)

Pour cette étude, nous avons rassemblé 313 verbes dénominaux issus du TLFi. Il s'agit de V préfixés en *a-, dé-, é-, et en/em-,* comme ABORDER, DÉFFRICHER, ÉCRÉMER, EMBOUTEILLER, et de V suffixés en *-iser* et *-ifier*, comme GLORIFIER, ÉTATISER. Nous avons annoté manuellement les noms bases selon le trait massif / comptable et les verbes dérivés selon le trait télique / atélique. Notre but est de vérifier empiriquement l'existence d'un lien entre la massivité du N et l'(a)télicité du V dérivé. Malgré les difficultés pratiques que ne manquent pas de générer des telles annotations, nous montrerons qu'il semble exister une corrélation entre ces caractéristiques puisqu'une majorité des V dénominaux téliques sont formés sur des N bases [+count]. Néanmoins, l'existence non marginale de V dénominaux atéliques formés sur des N bases [+count] devra être interrogée.

Exploitations morphosémantiques des embeddings lexicographiques

Basilio Calderone & Nabil Hathout, UMR CLLE-ERSS

Dans cet exposé, nous présentons plusieurs méthodes de construction d'espaces vectoriels sémantiques (ou *embeddings*) à partir des définitions de dictionnaire. Ces représentations sont créées au moyen de réseaux de neurones à partir des définitions du dictionnaire GLAWI. Ces embeddings peuvent notamment être construits en utilisant word2vec et une méthode basée sur les LSTM, capables de prendre en compte de manière plus fine l'ordre des mots. Nous verrons que les modèles construits en utilisant un bi-LSTM améliorent significativement les résultats. Ces méthodes permettent aussi de concevoir des systèmes capables de produire des définitions à partir de la forme d'un mot ou de prédire la catégorie « *unique beginners* » d'un mot à partir de sa définition.

Ces systèmes peuvent être utilisés pour extraire des classes de mots construits à partir de leurs définitions, pour trouver les mots qui correspondent à des définitions inédites ou pour associer des représentations sémantiques à des procédés dérivationnels (par exemple à la suffixation en *-able*).

Etudier les familles morphologiques à partir des lexèmes de basse fréquence. L'exemple de la dérivation verbale en français

Fabio Montermini & Juliette Thuilier, UMR CLLE-ERSS

Notre présentation se focalise sur la concurrence entre les procédés morphologiques permettant de créer un verbe à partir d'une base nominale ou une base adjectivale, en prenant en compte l'ensemble des procédés disponibles en français pour former ce type de dérivés (préfixation, suffixation, conversion). L'analyse proposée s'appuie sur un ensemble de verbes dénominaux et désadjectivaux ayant une fréquence 1 issus du corpus frWac. Ce jeu de données nous permettra de discuter les différents facteurs phonologiques, sémantiques et morphologiques qui organisent la

concurrence entre moyens dérivationnels. Une attention particulière sera portée au rôle que joue la famille dérivationnelle dans cette dynamique

Polysémie et troncation des noms en -ion en français

Rémi Anselme, Olivier Bonami & Heather Burnett, UMR LLF

Cette présentation rendra compte d'une étude qui reprend les hypothèses de Kerleroux (1999) sur l'apocope des noms en -ion en français. Après avoir montré qu'il existe des contre-exemples ponctuels à toutes les généralisations de Kerleroux, nous présenterons une étude de corpus qui suggère l'existence de contraintes non-catégoriques sur la troncation qui vont dans le sens de ces généralisations: il est faux que la troncation ne s'applique pas à des noms d'événements complexes, mais il est vrai que les formes tronquées ont plus rarement que leur base ce type d'acception.

Un résultat plus général et plus inattendu qui surgit de cette étude est le suivant: la troncation affecte la distribution des sens d'un nom de manière à rendre le sens plus prédictible, sans spécifiquement sélectionner un sens en particulier. Nous faisons l'hypothèse qu'il s'agit là d'une caractéristique générale des opérations dérivationnelles qui ne changent pas le domaine de dénotation de la base. Une étude pilote sur les verbes diminutifs suggère que cette hypothèse est sur la bonne voie.

Nous terminerons en discutant les conséquences de ces résultats pour les conceptions de la relation entre polysémie et morphologie constructionnels.

Références

Kerleroux, F. (1999). Sur quelles bases opère l'apocope? . Silexicales 2 : la morphologie des dérivés évaluatifs. D. Corbin, G. Dal, B. Fradin et al. Villeneuve d'Ascq, Presses de l'Université de Lille**:** 95-106.

Utilisation des frames pour la description paradigmatique des familles dérivationnelles

Daniele Sanacore, UMR CLLE-ERSS

La présentation montrera l'utilisation de structures inspirées des frames de FrameNet[1] pour essayer de trouver une réponse paradigmatique au problème de la représentation des relations sémantiques dans les familles dérivationnelles.

La sémantique des frames[2] et la ressource Framenet ont été pris comme point de départ. Le système d'annotation d'une famille dérivationnelle obéit à une structure à trois niveaux de représentation sémantique : argumentale, relationnelle et ontologique. En distinguant ces niveaux, nous pouvons superposer des frames associés à différentes familles et mettre ainsi en evidence des régularités de comportement sémantiques.

C'est ce que nous illustrerons à travers plusieurs exemples en français, après avoir présenté le cadre générale du travail, la base FrameNet et les raisons pour lesquelles cette ressource est insuffisante pour mener à bien notre objectif.

Références

[1] Baker, Collin F., Charles J. Fillmore, and John B. Lowe. "The berkeley framenet project." Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1998.

[2] Fillmore, Charles J. "Frame semantics and the nature of language." Annals of the New York Academy of Sciences 280.1 (1976): 20-32.

Prédiction semi-supervisée d'informations sémantiques

Cindy Aloui*, Lucie Barque°, Alexis Nasr* & Carlos Ramish*

* Laboratoire d'Informatique et Systèmes (LIS) et Université Aix Marseille, ° Laboratoire de Linguistique formelle (LLF) et Université Paris 13

Nous proposons dans cette étude une méthode minimalement supervisée pour prédire des informations sémantiques relevant du domaine nominal (l'opposition massif-comptable et l'opposition animé-inanimé) à partir de listes restreintes de noms non ambigus. Nous extrayons d'un large corpus brut l'ensemble des occurrences des noms des listes et entraînons un réseau de neurones à prédire la classe d'un nom à partir d'informations contextuelles. Le classifieur est ensuite utilisé pour prédire un score contextuel et estimer le score lexical (hors contexte) de noms inconnus en tirant profit des prédictions contextuelles les plus fiables. Notre méthode, évaluée sur deux jeux de données du français, atteint une précision de 90,1% pour l'opposition massif/comptable et 92,6% pour l'opposition animé/inanimé. Nous montrons que l'information sémantique acquise permet d'améliorer l'analyse syntaxique et la détection d'expressions multi-mots. Notre analyse de l'influence de chaque trait du modèle, de la taille de liste des graines et du traitement des noms polysémiques révèle qu'une liste de 100 noms par catégorie permet d'atteindre une précision raisonnable et que les scores contextuels sont en général ignorés au profit des scores lexicaux, excepté pour les noms polysémiques.