Journée ConSciLa

Dictionnaires et morphologie flexionnelle du français contemporain : état des lieux et perspectives

Vendredi 2 octobre 2015

Lieu : Université Paris 3-Sorbonne Nouvelle (salle à préciser)

coordonnée par A. GREZKA (CNRS LDI/Université Paris 13) et E. CARTIER (Université Paris 13/LIPN)

Comité d'organisation : E. Cartier (Université Paris 13/LIPN) ; W. Dekdouk (Université Paris 13/LDI) ; A. Grezka (CNRS LDI/Université Paris 13) ; C. Jacquet-Pfau (Collège de France/LDI) ; F. Martin-Berthet (Université Paris 13/LDI) ; M. Mathieu-Colas (Université Paris 13/LDI) ; J.-F. Sablayrolles (Université Paris 13/LDI)

Intervenants : E. Cartier (Université Paris 13/LIPN) ; L. Catach (Consultant indépendant et formateur) ; W. Dekdouk (Université Paris 13/LDI) ; A. Grezka (CNRS LDI/Université Paris 13) ; N. Hathout (CNRS CLLE-ERSS/Université Toulouse 2) ; C. Jacquet-Pfau (Collège de France/LDI) ; F. Martin-Berthet (Université Paris 13/LDI) ; M. Mathieu-Colas (Université Paris 13/LDI) ; F. Nemo (Université d'Orléans/LLL) ; J.-F. Sablayrolles (Université Paris 13/LDI) ; F. Sajous (CNRS CLLE-ERSS/Université Toulouse 2)

La Journée d'étude fera un état des lieux des dictionnaires électroniques (pour le TAL et pour l'édition électronique) du français contemporain, et notamment leur composante morphologique. Les défis et problèmes de leur exploitation dans le cadre du traitement automatique des langues, et de la connexion des grands corpus aux dictionnaires seront évoqués. La journée sera aussi l'occasion de présenter Morfetik (dictionnaire morphologique des mots simples du français du LDI) qui vient d'être mis à la disposition de la communauté. Enfin, une confrontation entre les dictionnaires pour le TAL et les dictionnaires éditoriaux sera lancée.

Les points suivants seront abordés :

- les dictionnaires électroniques pour le TAL du français contemporain
- confrontation dictionnaires/corpus/néologie/ terminologie ; interaction dictionnaire Morfetik avec bases de données de la DGLFLF
- confrontation dictionnaires électroniques/dictionnaires éditoriaux

PROGRAMME

09h30-09h40 : Ouverture et présentation de la journée
A. Grezka et E. Cartier

09h40-10h10 **Morfetik. Présentation et enjeux**
A. Grezka (CNRS LDI/Université Paris 13) ; M. Mathieu-Colas (Université Paris 13/LDI) ; F. Martin-Berthet (Université Paris 13/LDI)

10h10-10h50 **De GLÀFF à GLAWI, et retour… d'expérience sur six ans d'exploitation du Wiktionnaire**
F. Sajous (CNRS CLLE-ERSS/Université Toulouse 2)

10h50-11h30 **Morphonette et Démonette : deux ressources dérivationnelles morphologiques du français**
N. Hahtout (CNRS CLLE-ERSS/Université Toulouse 2)

11h30-11h50 : Pause

11h50-12h30 **Formes linguistiques, formes prosodiques, lexique et relations lexicales : vers des traitements automatiques**
F. Nemo (Université d'Orléans/LLL)

12h30-13h00 **Quelle place pour les « recommandations officielles » dans un dictionnaire ?**
C. Jacquet-Pfau (Collège de France/LDI)

13h00-14h30 : Déjeuner

14h30-15h10 **Les bases morphologiques dans les produits numériques du Robert**
L. Catach

15h10-15h40 **Morfetik et Neologia : deux bases en interaction**
J.-F. Sablayrolles (Université Paris 13/LDI) et W. Dekdouk (Université Paris 13/LDI)

15h40-16h00 : Pause

16h00-16h30 **Eléments et outils pour une confrontation continue dictionnaire - corpus**
E. Cartier (Université Paris 13)

16h30-17h00 **Discussion générale**

RESUMES DES COMMUNICATIONS

Morfetik. Présentation et enjeux

A. Grezka ; M. Mathieu-Colas ; F. Martin-Berthet

Le traitement automatique des langues exige avant toute chose une reconnaissance précise des formes, ce qui présuppose un recensement lexical aussi rigoureux et complet que possible. Dans cette perspective, nous avons entrepris d'élaborer une ressource linguistique ayant les caractéristiques suivantes : large couverture, précision et fiabilité des informations, respect des normes, évolutivité. Les données sont structurées sous forme de tables et servent de point de départ à un système de traitement (Morfetik) qui associe un moteur de flexion, un dictionnaire des formes fléchies, des interfaces de consultation et d'interrogation, ainsi qu'un ensemble d'outils permettant la maintenance et l'exploitation des ressources.
Cette ressource permet de générer automatiquement l'ensemble des formes simples du français : 758 035 formes fléchies (compte tenu des homographies), pour 102 962 lemmes. Elle est disponible dans un format XML et permet une consultation via un moteur de recherche. La donnée est accessible sous licence LGPL-LR : http://www-ldi.univ-paris13.fr/index.php?option=com_content&view=article&id=41&Itemid=41

De GLÀFF à GLAWI, et retour… d'expérience sur six ans d'exploitation du Wiktionnaire

F. Sajous

Cette communication fera une synthèse, sous forme de retour d'expérience, de six ans d'exploitation du Wiktionnaire pour le traitement automatique des langues et la description linguistique. Les potentiels de ce dictionnaire collaboratif seront notamment illustrés à travers la présentation de ressources telles que GLÀFF, un lexique flexionnel et phonologique, et GLAWI, un dictionnaire électronique du français au format XML, toutes deux conçues à partir du Wiktionnaire.

Morphonette et Démonette : deux ressources dérivationnelles morphologiques du français

N. Hathout

Les dictionnaires (électroniques) sont utilisés pour construire des ressources lexicales flexionnelles comme Morphalou ou GLÀFF. En revanche, peu de travaux portent sur la construction de lexiques dérivationnels. Deux exceptions sont notamment Morphonette, développé à partir des entrées de TLFnome en utilisant la mesure de similarité morphologiques Proxinette et l'analogie formelle, et Démonette, un réseau morphologique fortement redondant dans lequel les mots et les relations sont typés morpho-sémantiquement. La communication portera sur ces deux lexiques dérivationnels et sur quelques pistes permettant de les enrichir à partir d'un dictionnaire informatisé comme GLAWI.

Formes linguistiques, formes prosodiques, lexique et relations lexicales : vers des traitements automatiques

F. Nemo

Contexte

L'une des deux taches principales de la sémantique est de rendre compte de la diversité des emplois des signes (polysémie ou polycatégorialité), et ainsi de rendre compte de la génération du lexique. Cet objectif l'a depuis longtemps conduit à développer des modèles qui d'une façon ou d'une autre postulent - sous le nom notamment de signification ou encore d'instruction sémantique - l'existence de contraintes à satisfaire servant

d'input au processus d'interprétation et de différentes façons de satisfaire ces contraintes qui en sont les outputs et qui en se lexicalisant finissent par former le lexique. Pour l'essentiel, identifier de telles contraintes repose sur des techniques qui ne sont pas ou peu implémentables informatiquement. Les choses ont commencé à changer à partir du moment où au lieu de bloquer le signifiant pour étudier la diversité des sens, la sémantique linguistique a commencé à se poser la question de savoir quelle forme exactement codait les significations identifiées, et a découvert que les signifiants concernés n'étaient ni réductibles à leur seule forme phonématique, ni en ce qui concerne celle-ci, identifiables à des suites linéarisées de phonèmes, dans la mesure où à partir du début des années 2000, la forme prosodique des lexèmes polysémiques et le lien entre cette forme et leurs sens est devenu un objet d'étude à part entière, mais aussi dans la mesure où les unités porteuses de signification se sont révélées être fortement polymorphes (comme en témoigne minimalement la paire *forme*/*morph*). Car s'il est possible de repérer « à la main » les phénomènes concernés, prouver l'existence d'un lien entre forme prosodique et interprétation (lexicalisée) ou mesurer et décrire la polymorphie, suppose soit d'adopter des techniques de discrimination automatique soit d'étudier la polymorphie en disposant d'outils permettant un traitement exhaustif de l'ensemble du lexique.

Présentation

Dans ce contexte, l'objet de ma communication sera :

- de présenter des formats de représentations des mots qui intègrent la forme prosodique et les commentaires prosodiques qui leur sont associés, ainsi que la façon dont les entrées lexicographiques de dictionnaire électronique peuvent être restructurées pour rendre possible l'intégration de ces informations ;
- de présenter les méthodes, outils informatiques, techniques de classification et résultats du programme APR-IA DIASEMIE (Discrimination automatique du sens d'emploi des mots par l'intonation) mené à l'Université d'Orléans, et en particulier la chaîne de traitement qui va de l'extraction des données à la segmentation et vectorisation de leur forme, jusqu'à la phase (dite de navette) d'analyse des échecs de discrimination ;
- de présenter la démarche permettant de tester la polymorphie des formes phonétiques de l'ensemble du lexique français, et en particulier les formes d'alternances régulières ou d'adressage consonantiques identifiés ;
- d'objectiver et mesurer des formes non-décrites de structuration du lexique, à savoir l'existence de complexes lexicaux réunissant par des liens non-dérivationnels un ensemble de lexèmes associés à un morphème (ou une base morphémique) polymorphe.

Quelle place pour les « recommandations officielles » dans un dictionnaire ?

C. Jacquet-Pfau

Le dispositif d'enrichissement de la langue française, mis en place par décret du 3 juillet 1996 et modifié par décret du 25 mars 2015, a pour but de permettre un égal accès à l'information et aux savoirs et d'assurer la présence du français dans la vie sociale. Parmi les moyens mis en œuvre sous l'égide de la DGLFLF (Délégation générale à la langue française et aux langues de France), un dispositif interministériel permet de rendre régulièrement officiels des termes nouveaux destinés à enrichir la langue, de « combler les lacunes de notre vocabulaire et de désigner en français les concepts et réalités qui apparaissent sous des appellations étrangères ». Ce dispositif traite essentiellement, dans le cadre d'une procédure interministérielle, les termes étrangers et plus particulièrement les anglicismes qui sont très majoritairement utilisés pour désigner de nouveaux concepts ou de nouvelles réalités. Occasionnellement, il lui arrive de revenir sur des unités lexicales morphologiquement peu satisfaisantes et de préciser des définitions restées ambiguës et imprécises dans les domaines de spécialités.

Ces « équivalents », proposés par les différents groupes d'experts, sont soumis à la Commission d'enrichissement de la langue (parmi lesquels les « collèges d'experts ministériels » qui succèdent aux « commissions spécialisées de terminologie ») et, une fois validés par l'Académie française, sont publiés au *Journal officiel* et intégrés dans la base France Terme de la DGLFLF, consultable en libre accès par les internautes. Destinés à s'imposer dans les documents de travail des administrations et des établissements de l'État, ils sont également vivement recommandés auprès d'un public plus large et peuvent servir de référence, notamment pour les traducteurs et les rédacteurs techniques.

Il est donc nécessaire que la visibilité de ces nouveaux termes et leur diffusion soient aussi larges que possible. Quel outil serait plus directement adapté pour les faire reconnaître que les dictionnaires, la présence de ces nouveaux termes, ou du moins de certains d'entre eux, dans les dictionnaires étant par ailleurs un « marqueur » de leur bonne implantation dans l'usage courant ? Autre facette de l'attention à prêter à ces unités lexicales : leur repérage et leur interprétation dans des textes, officiels ou non, soit pour une utilisation ponctuelle et/ou spécialisée (rédaction, traduction interprétation, etc.), soit dans une application de traitement automatique (indexation, traduction, requête dans de grands corpus, etc.).
Nous proposons de nous interroger sur la nécessité de prendre en compte toutes ou certaines de ces « recommandations officielles » pour l'élaboration de la base lexicale de Morfetik et sur les modalités de cette initiative, notamment en examinant l'existant, autrement dit les pratiques dans quelques dictionnaires d'usage contemporain.

Les bases morphologiques dans les produits numériques du Robert

L. Catach

Les Dictionnaires Le Robert ont publié depuis 1986 des versions numériques de leurs principaux dictionnaires : le *Grand Robert de la langue française,* le *Petit Robert* (langue française et noms propres), le *Robert Junior,* etc., ainsi que les dictionnaires bilingues de la gamme *Robert & Collins.*
Toutes ces versions numériques utilisent des bases de données morphologiques du français (et parfois de l'anglais), qui sont utilisées à plusieurs niveaux dans les logiciels : module de conjugaison, affichage des féminins et pluriels, navigation hypertexte, recherches en texte intégral, etc.
Cet exposé explique la manière dont ces bases de données ont été constituées, les différentes questions liées à leur gestion et leur maintenance, ainsi que leur importance et leur mise en œuvre dans les produits numériques.

Morfetik et Neologia : deux bases en interaction

J.-F. Sablayrolles et W. Dekdouk

Une des voies pratiquées pour l'extraction automatique des néologismes formels est le recours à un corpus d'exclusion : tout mot d'un texte absent de ce corpus est un candidat néologisme. Candidat et pas automatiquement néologisme pour plusieurs raisons.

D'une part parce qu'il peut y avoir des fautes de frappe (inversion de lettres, lettres surnuméraires, fausses coupes…). Il faut noter par ailleurs que ce procédé non content de ne pas relever les néologismes sémantiques et les néologismes syntaxiques, laisse également passer les néologismes formels homonymiques d'un mot existant (*auditer*, dérivé inverse de *auditeur* homonyme de *auditer*, conversion du nom *audit*). Quoi qu'il en soit de ces lacunes, le problème principal demeure le choix de ce qui est pris comme corpus d'exclusion. De ce point de vue la base Morfetik est un des meilleurs corpus d'exclusion qui soient, tant par sa couverture : le grand nombre des items recensés (de par son mode de constitution puisant à diverses sources lexicographiques et encyclopédiques) que par la précision des informations flexionnelles. Ce qui est d'autant plus intéressant pour la base Neologia qu'à la différence de beaucoup d'autres centres s'occupant de veille néologique, nous considérons comme néologismes des innovations flexionnelles, autres que les simples fautes ou lapsus, du type *ils closirent*, *mon émolument*, etc.

D'autre part, les éléments relevés comme absents de la base sans qu'il y ait de fautes de frappe peuvent être des lacunes de la base Morfetik. Celle-ci, comme les autres, n'est pas complète et ne peut pas l'être. Peuvent en effet en être absents des mots rares, des mots anciens, des termes très spécialisés… ou encore et surtout des mots récents qui se sont diffusés mais qui ne sont pas encore intégrés dans les répertoires que dont les dictionnaires d'usage (la moyenne d'âge des mots entrés dans les dictionnaires est en effet élevée comme le montrent les calculs de Camille Martinez et aussi plusieurs études de Sablayrolles). Et c'est là où la base Neologia peut, en retour, être utile à Morfetik, en comblant certaines de ses lacunes, à condition de sélectionner dans ses néologismes ceux qui se sont suffisamment répandus pour mériter d'y être

intégrés. Il ne s'agit pas en effet de reverser automatiquement toutes nos prises, puisqu'il y a de nombreux hapax ou des créations de faible diffusion. Le suivi de diffusion automatique qui doit être ajouté à la base Neologia permettra de prendre des décisions en connaissance de cause. Mais là encore aucune décision mécanique ne peut être envisagée tant les paramètres à prendre en compte sont multiples.

Eléments et outils pour une confrontation continue dictionnaire - corpus

E. Cartier

Cette communication évoquera tout d'abord la nécessité d'une confrontation continue entre dictionnaire et corpus, le premier représentant la description de la langue, les corpus-discours permettant de prendre la mesure de la couverture des dictionnaires sur une période p, mais aussi de suivre l'évolution, à la fois du lexique mais aussi des usages grammaticaux de la langue. Nous partirons d'une expérience menée par (Sajous, 2014; Mathieu-Colas et al., 2015) fournissant une photographie de la couverture de différents dictionnaires existants (GLAFF, Morfetik, Lefff) sur plusieurs corpus (Wikipedia, dix ans de Le Monde, FrWacky). Nous évoquerons ensuite les pistes permettant le suivi synchronique des dictionnaires, au travers d'une plateforme en cours de mise en place dans le cadre du projet SPC Neoveille.

Références :

Sajous F., Hathout N., Calderone B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif. Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014), pp. 663-680, Berlin, Allemagne.

Grezka A., Cartier E., Mathieu-Colas M., (2015) Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL, Proceedings of TALN 2015, Caen.