Cahiers de l'ILSL-45 (2016 : 1-18)

Nouvelles technologies et standards méthodologiques en linguistique

Marianne KILANI-SCHOCH[1]
Christian SURCOUF[2]
Aris XANTHOS[3]
École de français langue étrangère (EFLE) (1, 2) & Section des Sciences du Langage et de l'Information (SLI) (1, 3), Université de Lausanne (CH)
Correspondance : marianne.kilanischoch@unil.ch

> L'informatique permet de renouer en morphologie avec la tradition philologique qui fondait la linguistique sur la lecture critique des textes. (HATHOUT, NAMER, PLÉNAT & TANGUY 2009)

1. Introduction

Dans le contexte toujours plus riche des Humanités numériques, ce numéro des Cahiers de l'ILSL présente les contributions à une journée d'étude consacrée aux nouvelles technologies et standards méthodologiques en linguistique qui s'est tenue en octobre 2014 à l'Université de Lausanne.

L'objectif de cette journée était de concentrer l'attention sur le volet linguistique des Humanités numériques pour en interroger la méthodologie. Les six contributions réunies ici explorent ainsi la question de la méthodologie dans des (sous)-disciplines linguistiques où l'empirie et les méthodes d'analyse ont été transformées par les nouvelles technologies. Cette thématique très générale abordée dans différents domaines linguistiques vise à évaluer les questions communes, et à travers elles à esquisser les directions vers lesquelles la linguistique tend à s'orienter aujourd'hui.

Un des premiers thèmes de réflexion est de toute évidence celui de l'extension des données empiriques. En rendant possible l'étude de collections toujours plus

importantes de données, dans des proportions que l'on n'avait pas soupçonnées auparavant, les nouvelles technologies étendent le champ et la portée de la recherche linguistique. Mais cet effet comporte le risque d'être surévalué et doit donc être étudié notamment pour éviter l'illusion que des données extensives se suffiraient à elles-mêmes. Car leur apport est naturellement fonction des objectifs linguistiques : sans hypothèses théoriques spécifiques, la linguistique de corpus et la linguistique quantitative demeurent purement descriptives.

La question est encore de savoir dans quelle mesure l'accès à un grand nombre de données tel qu'il est rendu possible par les nouvelles technologies influence et favorise certains choix théoriques (cf. VINCENT-DURROUX & CARR 2013). Par exemple, un objectif aussi ambitieux que celui de rendre compte de l'ensemble des comportements langagiers semble devenu concevable et caractérise même certaines versions radicales des grammaires basées sur l'usage et sur les exemplaires (voir PORT 2010, par exemple, et LADD 2011 pour une synthèse).

Cela revient à dire qu'aujourd'hui la question méthodologique des corpus et plus généralement celle de la linguistique de corpus ou basée sur des corpus (SCHEER 2013 : 19), en d'autres termes, la question de l'empirisme (VINCENT-DURROUX & CARR 2013 ; CHATER, CLARK, PERFORS & GOLDSMITH 2015) demeure une question épistémologique centrale. Avec elle, les conditions de validation des modèles linguistiques et la nature des généralisations constituent tout l'enjeu.

Dans la linguistique de corpus, on peut distinguer les problèmes méthodologiques relatifs aux corpus eux-mêmes, qu'il s'agisse des limites inhérentes à l'échantillonnage ou de celles qui tiennent à leur utilisation et aux outils de description, et les problèmes méthodologiques spécifiques aux instruments de traitement et de modélisation. Ces problèmes sont à envisager dans leurs conséquences sur la

nature des généralisations linguistiques (ERNESTUS & BAAYEN 2011 : 388). Nous verrons que les contributions des conférenciers s'inscrivent toutes dans l'une ou l'autre de ces problématiques.

1.1. Recueil des ressources linguistiques, construction des corpus

Les problèmes classiques de la constitution d'un corpus, à savoir la pertinence quantitative (quel seuil d'occurrence ? DAL & NAMER 2012 : 1265), la représentativité et l'équilibre (entre des données de différents genres, registres, par exemple), l'homogénéité et l'exhaustivité (TOMASELLO & STAHL 2004) se posent-ils différemment aujourd'hui ? La question de savoir quel volume de données est suffisant (TOMASELLO & STAHL 2004) et quelle stratégie d'échantillonnage (*sampling*) est la plus appropriée n'a évidemment pas disparu avec l'accès à des données toujours plus nombreuses. Elle a pris un nouveau relief avec l'avènement des mégadonnées (*big data*). Mais la réflexion sur les limites des corpus ainsi que sur les méthodes pour en amoindrir l'impact s'est également développée (BAAYEN 2008 ; MALVERN, RICHARDS, CHIPERE, & DURÁN 2004 ; XANTHOS & GILLIS 2010).

Du côté d'une perspective critique, SCHEER (2013) par exemple, dénonce l'illusion selon laquelle le corpus produirait la science par lui-même ou même garantirait le statut scientifique de l'analyse menée (VINCENT-DURROUX & CARR 2013). SCHEER considère que le corpus n'est pas plus pertinent aujourd'hui qu'il ne l'était hier même si « meilleures » sont les données, plus grandes sont les possibilités de réfuter les théories. Il rappelle le rôle déterminant de la conception et de la construction du corpus (*corpus design*) et des finalités théoriques qu'il sert.

TOMASELLO & STAHL (2004 : 119) avaient déjà montré, pour le domaine de l'acquisition, que les procédures

d'échantillonnage, par exemple, dépendent des questions de recherche et ne peuvent être universelles. On observe aujourd'hui des approches diverses et complémentaires dans le recours aux données linguistiques empiriques et la construction des corpus. D'une part la recherche et l'exigence de données multiformes et de mégadonnées – renforçant par exemple le potentiel statistique d'un corpus mais posant toutes sortes de problèmes techniques – connaissent un grand essor (notamment en psycholinguistique, voir KEULEERS & BALOTA 2015 ; TSUJI 2014 ; ROY, FRANK & ROY 2014 ; NEWMAN, RATNER & ROWE 2014). Il ne semble plus y avoir de limite supérieure à ce que peut être un corpus (VAUGHAN & CLANCY 2013), « toujours augmentable et jamais fini » selon la perspective de SINCLAIR (1991) (ARBACH & ALI 2013 : 23).

D'autre part, la nécessité de corpus plus petits, facilement annotables, accessibles et adaptés à des questions spécifiques, c'est-à-dire des corpus sur mesure, refait surface (KOESTER 2010 ; VAUGHAN & CLANCY 2013 ; ROMERO-TRILLO 2013). Elle soulève la question de la pertinence de la recherche empirique basée sur des corpus informatisés selon les domaines ou sous-disciplines linguistiques : les problèmes techniques rencontrés dans certains d'entre eux peuvent constituer une restriction sévère (SCHEER 2013).

Par exemple, les problématiques phonologiques ou morphologiques sont généralement plus faciles à étudier empiriquement que les problématiques pragmatiques, syntaxiques ou prosodiques. En pragmatique les recherches informatisées sont rendues difficiles, par exemple, par le fait qu'il est particulièrement laborieux d'annoter/coder les actes de langage, et parce que les phénomènes sont par définition dépendants du contexte (VAUGHAN & CLANCY 2013). Plusieurs publications récentes plaident en faveur de petits corpus pragmatiques contextualisés, relevant que la pertinence de ceux-ci est un effet secondaire des développements

technologiques qui facilitent l'accessibilité des grands corpus et rendent le choix possible (VAUGHAN & CLANCY 2013, ROMERO-TRILLO 2013). Ces petits corpus permettent l'accès à des métadonnées complètes (voir 1.2).

Et la voie moyenne consistant à adopter des corpus de référence, c'est-à-dire de larges ensembles de données sans qu'il s'agisse de mégadonnées[1], rigoureusement construits et répondant aux exigences de différents descripteurs, est une autre option, nécessaire pour un « usage responsable des données de corpus », dans la recherche sur la variation en phonologie, par exemple (ERNESTUS & BAAYEN 2011 : 384).

1.2. Utilisation des ressources : métadonnées

La méthodologie appliquée dans l'utilisation des ressources électroniques satisfait-elle aux exigences de rigueur établies par une longue tradition en sciences du langage, qu'il s'agisse de la tradition philologique ou de la tradition sociolinguistique, avec l'objet desquelles les données et corpus d'internet peuvent être comparés ? On constate le plus souvent que les métadonnées, c'est-à-dire l'ensemble standard de descripteurs de données permettant de caractériser les ressources digitales (BURNARD 2005 : 40), à savoir les conditions de constitution des corpus ou du recueil des données, l'information sur les locuteurs/scripteurs, les éléments de contextualisation, le statut discursif, diatopique, diastratique, diaphasique, diamésique et sociolinguistique en général, sont inexistantes. Or, privé de ces métadonnées un corpus n'est qu'une collection décontextualisée de mots (BURNARD 2005 : 41 ; ADOLPHS & KNIGHT 2010) sans représentativité possible (ARBACH & ALI 2013 : 17).

Il faut relever par ailleurs qu'on ne dispose pas d'un inventaire standard des outils descriptifs et analytiques (par

[1] La question étant bien entendu celle des critères de distinction entre les deux types de données.

ex. les différents types de mesures) disponibles pour interroger les ressources digitales alors que des initiatives ont été prises il y a plus de 30 ans pour standardiser la création et la gestion de données linguistiques et textuelles digitales (par exemple TEI = Text Encoding Initiative, BURNARD 2014). Le projet de *General Ontology for Linguistic Description* (*GOLD*) discuté par FARRAR 2013, à savoir une base partagée de connaissances linguistiques destinées au traitement informatique (*machine processing*) relève probablement de la même préoccupation du partage des ressources et des instruments.

Un autre problème méthodologique tient au fait que la diversification des sources, leur comparabilité relativement à la sélection et à la taille des corpus/échantillons ou à la fréquence (relative ou absolue) des unités, par exemple, le codage des corpus, leur annotation (souvent manquante, même au niveau morphosyntaxique) et le contrôle de la qualité de ces traitements, sont peu problématisés. De même, la différenciation entre divers types de données informatisées, par exemple les données de la Toile et les données de corpus, ne connait pas encore de critères établis (DAL & NAMER 2012). Pourtant, la question de la quantité et de la variété des sources informatisées devant produire le même résultat pour qu'une généralisation soit considérée comme fiable fait partie des problèmes élémentaires de validation.

L'accessibilité de sources informatisées variées soulève ensuite la question des conséquences du caractère de plus en plus interdisciplinaire de la recherche linguistique sur la nature des indices ou éléments de preuve. Dans quelle mesure et jusqu'à quel point les résultats des autres disciplines et sous-disciplines doivent-ils être pris en considération dans le système de validation ? (ERNESTUS & BAAYEN 2011 : 389 ; DRESSLER 2013).

2. Outils de traitement et de modélisation

Le problème de la méthodologie, et celui des conditions de validation à l'aide de ces données, concerne aussi bien les ressources fournies par les nouvelles technologies que les outils de traitement utilisés.

Certaines sous-disciplines de la linguistique, telle la psycholinguistique de l'acquisition dont l'approche a été entièrement renouvelée par les technologies numériques, mettent la méthodologie quantitative au centre de la recherche et de l'argumentation (voir par exemple SAFFRAN, ASLIN & NEWPORT 1996 ; SAFFRAN, NEWPORT & ASLIN 1996 ; REDINGTON, CHATER & FINCH 1998 ; ASLIN, SAFFRAN & NEWPORT 1999 ; LEWIS & ELMAN 2001 ; BOD 2009 ; LIEVEN 2014).

On peut se demander si ces sous-disciplines préfigurent le devenir de la linguistique, voire des sciences humaines dans leur ensemble et si un nouveau standard méthodologique est en train d'émerger rendant potentiellement caduques les recherches conduites sans outillage informatique.

On notera dans ce sens que le tournant quantitatif à la fin du siècle passé et le développement des modèles basés sur l'usage et les exemplaires semblent avoir réorienté les problèmes de modélisation linguistique du côté de la nature de la modélisation informatique. Car, comme le rappellent DAL & NAMER :

> depuis une dizaine d'années, [...] en quelque sorte, l'usage est à la portée du linguiste, grâce d'une part à la démocratisation spectaculaire des capacités de stockage des ordinateurs de plus en plus performantes, et d'autre part à l'évolution des techniques informatiques de recherche d'information, qui simplifient et accélèrent la fouille de ces grandes quantités de données informatisées. (DAL & NAMER 2012 : 1264)

Dans une certaine mesure, ce développement a remis sur le devant de la scène le débat ancien concernant l'opposition langue-parole/compétence-performance et interroge l'objet

ultime de la linguistique. Les outils technologiques fournissent potentiellement aux théories linguistiques les moyens de rendre compte non seulement des systèmes linguistiques dans leur complexité mais encore des actes de parole individuels (LADD 2011) dans la double perspective du locuteur et de l'interlocuteur, c'est-à-dire en intégrant à la fois la production et la compréhension (ERNESTUS 2014). Il convient de se demander dans quelle mesure les modèles linguistiques devront désormais chercher à intégrer toute la diversité des usages et donc d'élaborer des outils informatiques à même de rendre compte de la nature variée des dimensions qui y sont impliquées (ERNESTUS & BAAYEN 2011 ; ERNESTUS 2014). Ceux-ci auront à traiter à la fois les probabilités spécifiques à certains items, l'analogie dynamique (*dynamic analogy-driven computation*) et la compression des données (BAAYEN 2007 : 98 ; ERNESTUS & BAAYEN 2011).

L'intérêt croissant, cette dernière décennie, pour la complexité des systèmes linguistiques notamment en morphologie (ALBRIGHT & HAYES 2002 ; BAAYEN 2007 ; BAAYEN, MILIN, ĐURKĐEVIĆ, HENDRIX & MARELLI 2011 ; BAERMAN, BROWN & CORBETT 2015) est un autre exemple de recherches récentes soutenues par les nouvelles technologies et représentant un défi informatique aussi bien que linguistique.

En somme, la question fondamentale suscitée par les nouvelles technologies au niveau des données, de leur traitement ou de la modélisation est celle de savoir si elles représentent un défi pour les modèles linguistiques antérieurs sur le plan de l'objet comme sur celui des résultats acquis. Dans quelle mesure remettent-elles en question la pertinence de ces modèles et jettent-elles le discrédit sur les généralisations qui ont été énoncées ?

Ce sont les approches connexionnistes, il y a quelques décennies, qui ont inauguré le débat sur la validité des

catégories et unités linguistiques fondamentales telles que le phonème, le morphème, le mot-forme, etc. et sur la plus grande pertinence des données statistiques et des architectures subsymboliques (cf. BAAYEN *et al.* 2011 ; PORT 2010, etc.).

Aujourd'hui, en morphologie, pour ne prendre que cet exemple, des chercheurs comme Harald BAAYEN, Gilles BOYÉ, Olivier BONAMI, parmi d'autres, font l'impasse sur le concept de règle, lui préférant celui de généralisations probabilistes qui peuvent être obtenues en recourant à des techniques statistiques et d'apprentissage automatique (*machine learning*). Certaines versions de la théorie par exemplaires vont jusqu'à rejeter l'existence de sous-disciplines linguistiques comme la phonologie (voir LADD 2011 : 368).

Est-il ainsi plausible que, pour suivre SCHEER (2013 : 1, 4), toute proposition scientifique doive dorénavant être statistiquement pertinente ? Une manière plus spécifique et constructive de formuler la question consiste à se demander, à la suite d'ERNESTUS & BAAYEN (2011 : 387), « comment des analyses statistiques basées sur des corpus sont articulées avec la théorie de la grammaire ».

Parallèlement à ces débats, il y a à s'interroger sur les conséquences de ces évolutions pour le reste de la discipline. Les prochaines années diront la place réservée aux recherches qui ne recourent pas aux techniques informatiques. Elles montreront si celles-ci deviennent simplement obsolètes.

Les auteurs qui ont contribué à ce numéro ne prétendent ni ne peuvent évidemment répondre à l'ensemble de ces questions. Mais chacun apporte une réflexion particulière sur les possibilités et/ou les limites de l'apport des nouvelles technologies. Leurs travaux portent sur les divers domaines de l'acquisition du langage, du traitement du langage, de la phonologie, de la morphologie flexionnelle, dérivationnelle et computationnelle. Ils partagent une vaste expérience en

matière de bases de données et de corpus de grandes dimensions, une connaissance approfondie des nouvelles technologies et l'exigence d'une méthodologie solide.

Dans ce numéro nous proposons la traduction française des conférences de cette journée tenues en anglais ainsi que celle d'une contribution allemande de DRESSLER, KORECKY-KRÖLL & MÖRTH dont la participation à la journée n'a pas été possible. À l'exception de l'article de Nabil HATHOUT, qui a rédigé en français la version écrite de sa conférence, et de celui de DRESSLER, KORECKY-KRÖLL & MÖRTH, traduit par Marianne Kilani-Schoch, les textes de ce numéro ne sont donc pas des articles préparés par les auteurs eux-mêmes mais des traductions de leur conférence du 17 octobre 2014, élaborées par nos soins à partir de la transcription des enregistrements de la Journée effectuée par Guillaume Feigenwinter et supervisée par Marianne Kilani-Schoch, et d'une première traduction de Guillaume Feigenwinter que Marianne Kilani-Schoch, Christian Surcouf et Aris Xanthos ont adaptée et reformulée. Nous avons pris le parti de conserver certaines caractéristiques de la présentation orale, qu'il s'agisse du style, de l'adresse à l'audience, des renvois aux autres conférences ou du jeu de questions-réponses en fin de texte.

3. Présentation des contributions

La contribution de BOYÉ aborde une série de problèmes méthodologiques liés à l'étude de la morphologie flexionnelle et en particulier la question du remplissage des paradigmes (ACKERMAN *et al.* 2009) dans la conjugaison des verbes en français. Dans une première partie, différents aspects problématiques sont définis : (i) au niveau des formes, les représentations phonologiques, la variation des formes et l'influence des fréquences sur le lexique, (ii) au niveau des cases, la surabondance et la défectivité. Un tour d'horizon des données disponibles tend à montrer que la prise en

compte de ces obstacles n'est pas chose aisée pour l'instant. La deuxième partie présente l'ébauche d'un modèle à même de contourner ces difficultés sur la base d'un lexique d'entrainement prenant en compte des fréquences, des représentations variées, la surabondance (sans inclure la défectivité). L'analyse est basée sur la collecte des analogies entre formes dans l'échantillon lexical et des classes de compétition entre analogies pour produire en masse des formes-candidates. Le paradigme flexionnel de chaque lexème est ensuite extrait en choisissant une clique à couverture maximale parmi l'ensemble des formes produites.

La contribution de DRESSLER, KORECKY-KRÖLL & MÖRTH s'attache à la question de la prévisibilité et de la prédictibilité probabiliste en linguistique, ainsi qu'au rôle des corpus électroniques dans l'élaboration des prévisions et de la vérification empirique, en se concentrant plus particulièrement sur deux domaines : l'acquisition de la langue première et le développement diachronique.

Les auteurs détaillent la variété des facteurs à prendre en considération dans le premier domaine pour rendre compte de sa complexité et effectuer des prédictions relatives à l'acquisition des pluriels allemands, turcs et anglais, notamment, qu'il s'agisse des facteurs typologiques de richesse, transparence et univocité morphologique de la flexion de la langue-cible, de la relation entre intégration et production chez l'enfant ou du niveau socioéconomique des familles. Les prédictions probabilistes sont limitées à la richesse de l'input et de l'output et à la relation entre les deux dimensions pour lesquelles on dispose de mesures quantifiables.

Les rétrodictions de la diachronie sont également discutées à l'aune d'un exemple de développement morphologique dans les dialectes italo-romans qui a pu faire l'objet de rétrodictions relativement précises. Les auteurs

montrent pourquoi une telle prévisibilité demeure exceptionnelle en diachronie.

Dans sa contribution, ERNESTUS évoque l'intérêt des grands corpus oraux dans la recherche en linguistique et en psycholinguistique, par le fait même qu'ils permettent de dépasser les limites de l'expérimentation en laboratoire ou du recours à l'intuition par le linguiste. Ainsi dans un premier temps, sur la base d'une analyse quantitative du Corpus Oral du Néerlandais, l'auteure démontre que l'assimilation régressive dans cette langue ne fonctionne pas toujours conformément à la description proposée par les linguistes. Si le recours à des corpus oraux de grande taille permet de déterminer la structure phonétique des mots dans le flux effectif de la parole, ERNESTUS rappelle néanmoins que leur utilisation doit s'accompagner de certaines précautions. Elle souligne plus particulièrement les difficultés de la transcription phonétique, qu'elle soit réalisée manuellement par des transcripteurs humains ou automatiquement à l'aide d'un dispositif informatique de reconnaissance vocale, qu'il faudra de toute façon alimenter correctement pour l'obtention de résultats exploitables. L'auteure aborde par la suite les écueils du traitement statistique des données, notamment celui de la manière d'intégrer les prédicteurs en fonction de leur nombre et de leur degré de corrélation. À la suite de ces divers rappels sur les précautions nécessaires à la manipulation des corpus oraux, ERNESTUS conclut que la recherche en linguistique et en psycholinguistique ne peut se dispenser de leur apport dans la mesure où les corpus oraux, contrairement aux expérimentations en laboratoire, donnent accès à ce que font *réellement* les locuteurs dans leur pratique quotidienne de l'oral.

La contribution de GILLIS dresse un panorama critique de l'usage des technologies pour l'étude de l'acquisition – en particulier dans son volet observationnel, centré sur les

corpus oraux. Partant du constat que la manière dont de tels corpus sont traditionnellement constitués est extrêmement couteuse en temps de travail, alors même qu'une proportion très restreinte des productions enfantines est échantillonnée, l'auteur examine successivement plusieurs façons de remédier au problème de la rareté des données : la base de données CHILDES, le système LENA™ et l'approche « big data » mise en place par Deb ROY (2011). Ce passage en revue suggère que le problème ne se limite pas à la difficulté d'enregistrer une portion représentative de l'input et de l'output enfantin ; il faut encore et surtout pouvoir transcrire et annoter les données récoltées, ce que GILLIS conçoit comme un défi majeur dans la perspective d'une avancée substantielle de ce domaine de recherche.

L'article de HATHOUT présente en détail les avantages et les couts de l'évolution consécutive au développement des nouvelles technologies pour la recherche en morphologie, et soulève la question de la nature et de la place des données dans cette recherche.

Ce qu'on appelle désormais la morphologie extensive, fondée sur de vastes quantités de données, a connu déjà différentes périodes dans une histoire que HATHOUT retrace en quelques pages et qui va de l'accès très large à la Toile dans les années 90 aux restrictions que les moteurs de recherche ont imposées aujourd'hui par la protection des index.

Si le chercheur montre, à l'exemple des travaux qu'il a menés en équipe sur les adjectifs en *-esque* et en *-able*, que la quantité de données prises en compte détermine directement la qualité des résultats, il expose aussi la longue liste des problèmes méthodologiques posés par les données de la Toile. Il évoque en outre les transformations que l'approche extensive engendre dans la recherche en morphologie devenue expérimentale, pour appeler à une

revalorisation du travail de constitution de ressources et de collections de données morphologiques en même temps qu'à une politique de partage dont il esquisse les contours.

TRIBUSHININA & MAK abordent la question des limites de l'étude de corpus, donc, de la production verbale dans l'analyse du développement linguistique chez des enfants bilingues et des enfants atteints d'un trouble du langage.

Ils montrent à travers leur recherche sur les connecteurs et les pronoms sujets en russe et en néerlandais que la différenciation linguistique des deux populations, souvent confondues au niveau de la production, nécessite le recours au paradigme méthodologique du monde visuel et à la technique de l'oculométrie (*eye-tracking*) : dans les deux domaines de la langue étudiés, l'écart significatif de compétence linguistique entre enfants bilingues et enfants atteints d'un trouble spécifique du langage (TSL/SLI) n'apparait qu'avec les mesures précises de l'activité de traitement rendues possibles par l'oculométrie.

Nos remerciements vont à François Rosset (doyen de la Faculté des lettres en 2014) et Thérèse Jeanneret (directrice de l'EFLE) pour leur soutien financier dans l'organisation de la Journée ainsi qu'au Centre de linguistique et des sciences du langage (CLSL) et à son directeur, Marcel Burger, pour les fonds accordés en vue de cette publication.

Références

ACKERMAN Farrell, BLEVINS James P. & MALOUF Robert (2009). Parts and Wholes: Implicative Patterns in Inflectional Paradigms. In BLEVINS James P. & BLEVINS Juliette (Eds), *Analogy in Grammar: Form and Acquisition*. Oxford: Oxford University Press, 54-82.

ADOLPHS Svenja & KNIGHT Dawn (2010). Building a Spoken Corpus: what are the Basics? In O'KEEFFE Anne & MCCARTHY Michael (Eds), *The Routledge Handbook of Corpus Linguistics.* Abingdom: Routledge, 38-52.

ALBRIGHT Adam & HAYES Bruce (2002), Modeling English Past Tense Intuitions with Minimal Generalization. In MAXWELL Michael (Ed.), *Proceedings of the sixth Meeting of the ACL Special Interest Group in Computational Phonology*. Philadelphia: ACL, 58-69.

ARBACH Najib & SAANDIA Ali. (2013). Aspects Théoriques et Méthodologiques de la Représentativité des Corpus. *Corela-HS-13.* Publié en ligne le 10.12.2013.

ASLIN Richard N., SAFFRAN Jenny R. & NEWPORT Elissa L. (1999). Statistical Learning in Linguistic and Nonlinguistic Domains. In MACWHINNEY Brian (Ed.), *The Emergence of Language.* Mahwah, NJ: Lawrence Erlbaum, 359-380.

BAAYEN R. Harald. (2007). Storage and Computation in the Mental Lexicon. In JAREMA G. & LIBBEN G. (Eds), *The Mental Lexicon*. Amsterdam: Elsevier, 81-104.

BAAYEN R. Harald. (2008). *Analyzing Linguistic Data*. Cambridge: Cambridge University Press.

BAAYEN R. Harald, MILIN Petar, ĐURĐEVIĆ Dusica Filipović, HENDRIX Peter & MARELLI Marco (2011), An Amorphous Model for Morphological Processing in Visual Comprehension based on Naive Discriminative Learning, *Review 118-3*, 438–482.

BAERMAN Matthew, BROWN Dunstan & CORBETT Greville G. (Eds) (2015). *Understanding and Measuring Complexity*. Oxford: Oxford University Press.

BOD Rens (2009). From Exemplar to Grammar: a Probabilistic Analogy-Based Model of Language Learning. *Cognitive Science* 33, 752-793.

BURNARD Lou (2005), Metadata for Corpus Work. In WYNNE Martin (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Chapter 3. Produced by ahds literature, language and linguistics. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

BURNARD Lou (2014). *What is the Text Encoding Initiative?* OpenEdition Press.

CHATER Nick, CLARK Alexander, PERFORS Amy & GOLDSMITH John A. (2015). *Empiricism and Language Learnability*. Oxford: Oxford University Press.

DAL Georgette & NAMER Fiametta (2012). Faut-il Bruler les Dictionnaires ? Ou comment les Ressources Numériques ont Révolutionné les Recherches en Morphologie. *Congrès Mondial de Linguistique Française*, 1261-1276. EDP Sciences. DOI10.1051/shsconf/20120100217.

DRESSLER Wolfgang U. (2013). Quo vadis linguistica ? Conférence plénière. Salzburg. 40. *Oesterreichische Linguistiktagung*. Ms.

ERNESTUS Mirjam (2014). Acoustic Reduction and the Roles of Abstractions and Exemplars in Speech Processing. *Lingua 142*, 27-41.

ERNESTUS Mirjam & BAAYEN, R. Harald (2011). Corpora and Exemplars in Phonology. In GOLDSMITH John, RIGGLE Jason & YU Alan C.L., *The Handbook of Phonological Theory.* Malden, Oxford: Wiley-Blackwell, 374-400.

FARRAR Scott (2013). An Ontological Approach to Canonical Typology: Laying the foundations for e-Linguistics. In BROWN Dunstan, CHUMAKINA Marina & CORBETT Greville G., *Canonical Morphology and Syntax.* Oxford: Oxford University Press, 239-261.

HATHOUT Nabil, NAMER Fiametta, PLÉNAT Marc & TANGUY Ludovic (2009). La Collecte et l'Utilisation des Données en Morphologie. In FRADIN Bernard, KERLEROUX Françoise & PLÉNAT Marc (ss la dir.), *Aperçus de Morphologie du Français*. Paris : Presses Universitaires de Vincennes, 267-289.

KEULEERS Emmanuel & BALOTA David A. (2015). Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview of Recent Developments. *Quarterly Journal of Experimental Psychology 68-8*, 1457-1468.

KOESTER Almut (2010). Building Small Specialized Corpora. In O'KEEFFE Anne & MCCARTHY Michael (Eds), *The Routledge Handbook of Corpus Linguistics.* Abingdom: Routledge, 66-79.

LADD, D. Robert (2011) Phonetics in Phonology. In GOLDSMITH John, RIGGLE Jason & YU Alan C.L., *The Handbook of Phonological Theory.* Wiley-Blackwell, 348-373.

LEWIS John D. & ELMAN Jeffrey L. (2001). Learnability and the Statistical Structure of Language: Poverty of Stimulus Arguments Revisited. *Proceedings of the Twenty-Sixth Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press, 359-370.

LIEVEN Elena (2014) First Language Development: a Usage-Based Perspective on Past and Current Research. *Journal of Child Lang*uage *41-Supplement S1*, 48-63.

MALVERN David D., RICHARDS Brian J., CHIPERE Ngoni & DURÁN Pilar. (2004). *Lexical Diversity and Language Development.* Basingstoke: Palgrave Macmillan.

NEWMAN Rochelle, RATNER Nan & ROWE Meredith (2014). Big Data: Challenges of Conducting Longitudinal Studies. Amsterdam. *IASCL Symposium*. Abstracts, 324.

PORT Robert F. (2010). Rich Memory and Distributed Phonology. *Language Sciences* 32, 43-55.

REDINGTON Martin, CHATER Nick & FINCH Steven. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science 22-4*, 425-469.

ROMERO-TRILLO Jesús. (2013). *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*.

ROY Brandon C., FRANK Michael C. & ROY Deb (2014). Harnessing Big Data in a Naturalistic Study of one Child's Early Word Learning. Amsterdam. *IASCL Symposium*. Abstracts, 323.

ROY Deb (2011), The Birth of a Word, http://www.ted.com/talks/deb_roy_the_birth_of_a_word, [22/08/2015].

SAFFRAN Jenny R., ASLIN Richard N. & NEWPORT Elissa L. (1996), Statistical Learning by 8-Month-Old Infants, *Science 274-5294*, 1926-1928.

SAFFRAN Jenny R., NEWPORT Elissa L. & ASLIN Richard N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language 35-4*, 606-621.

SCHEER Tobias (2013). The Corpus: a Tool among Others. *Corela-HS-13.* Publié en ligne le 25.11.2013.

SINCLAIR, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

TOMASELLO Michael & STAHL Daniel. (2004). Sampling Children's Spontaneous Speech: How Much is Enough? *Journal of Child Language 31-01*, 101-121.

TSUJI Sho. (2014). Big Data in Infant Language Acquisition. Chances and Challenges. Amsterdam. *IASCL Symposium*, Absracts, 322.

VAUGHAN Elaine & CLANCY Brian (2013). Small Corpora and Pragmatics. *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies.* Dordrecht: Springer, 53-76.

VINCENT-DURROUX Laurence & CARR Philippe (2013). Statut et Utilisation des Corpus en Linguistique. *Corela-HS-13.* Publié en ligne le 11.12.2013.

XANTHOS Aris & GILLIS Steven. (2010). Quantifying the Development of Inflectional Diversity. *First Language 30-2*, 175-198.