JéTou 2017

Les Interfaces en Sciences du Langage

Interfaces in Linguistics

Actes des Journées d'études toulousaines 2017
18 et 19 mai 2017
Université Toulouse – Jean Jaurès

Conférencières invitées/Keynote speakers

Marie Lallier

Développement de la lecture et bilinguisme précoce

Je présenterai des données comportementales et neurophysiologiques obtenues chez des bilingues precoces qui ont appris/apprennent à lire simultanément dans deux langues variant ou non en terme de transparence orthographique et système phonologique. Nous montrons que des interactions cross-linguistiques ont lieu durant l'apprentissage de la lecture chez les bilingues, qui influencent l'utilisation de certaines strategies de lecture (sub-lexicale, lexicale). Nous concluons que le développement de la lecture chez les bilingues est en partie déterminé par le degré de similarité entre les langues sur leur transparence orthographique et leur répertoire phonologique. Ces résultats ont des implications pour le diagnostique des troubles de lecture chez les individus bilingues.

Audrey Bürki

Interface oral/écrit, ou le rôle du langage écrit dans la production et la reconnaissance des mots

De nombreuses études psycholinguistiques ont mis en évidence le rôle des connaissances orthographiques sur les performances des participants dans diverses tâches de reconnaissance (et dans une moindre mesure, de production) de la parole. Les mécanismes cognitifs permettant d'expliquer cette influence sont sujets à débats. La variation phonologique offre un moyen de choix pour l'étude de cette question. Dans cette présentation, je détaillerai les résultats de plusieurs études ayant examiné le rôle de l'orthographe à travers la variation phonologique. Les résultats de ces études seront discutés à la lumière des études précédentes sur le rôle de l'orthographe, et des modèles psycholinguistiques de production et reconnaissance des mots parlés.

Table des matières / *Table of Contents*

Organisation des JéTou / JéTou organisation

Actes des JéTou / Proceedings of JéTou

Session communications orales 1 / Oral session 1

Session communications orales 2 / Oral Session 2

Session communications affichées 1 / Poster session 1

Session communications orales 3 / Oral session 3

Session communications orales 4 / Oral session 4

Session communications affichées 2 / Poster session 2

Appel à communications

Les doctorantes de deux laboratoires de Sciences du Langage de l'Université de Toulouse:

- CLLE-ERSS (Équipe de Recherche en Syntaxe et Sémantique)
- Octogone-Lordat (Laboratoire de neuropsycholinguistique)

organisent la 6e édition des **JéTou** (*Journées d'études Toulousaines*).

Ces journées s'adressent aux étudiants en Master, aux doctorants et aux jeunes chercheurs (jusqu'à trois ans après la soutenance) en Sciences du langage.

Les Sciences du Langage (SDL) deviennent un domaine de recherche de plus en plus interdisciplinaire. Ceci n'est pas surprenant étant donné la nature du langage lui-même : différents niveaux de la structure linguistique sont en continuelle interaction, et le langage a des interfaces avec de nombreuses activités. L'objectif du colloque JéTou 2017 est de réunir les jeunes chercheurs qui travaillent sur *différents types d'interdisciplinarité au sein des SDL, mais aussi entre les SDL et d'autres disciplines scientifiques*. Nous accueillons donc tous travaux intégrant deux ou plusieurs disciplines scientifiques afin de répondre à une question linguistique. Ces travaux peuvent explorer les interfaces entre les différents niveaux de description linguistique théorique (phonétique, phonologie, morphologie, syntaxe, sémantique, discours), l'interaction de la linguistique théorique avec d'autres disciplines des SDL (acquisition et apprentissage, enseignement du langage, traductologie, sémiologie, etc.), ou entre les SDL en général et d'autres disciplines scientifiques comme la psychologie, la neurologie, la sociologie, l'anthropologie, l'informatique, etc.

Les thématiques centrales du colloque seront les interfaces entre 1) la linguistique et l'informatique, 2) la linguistique, la psychologie et la neurologie, 3) la linguistique, l'acquisition, l'apprentissage et l'enseignement des langues ; cependant, toute proposition de nature interdisciplinaire sera considérée à titre égal. Les travaux articulant réflexions théoriques et données attestées seront particulièrement appréciés.

Liste non-exhaustive des domaines explorés dans les papiers :

- *Acquisition du langage*
- *Apprentissage du langage*
- *Cognition*
- *Discours*
- *Enseignement des langues*
- *Lexicographie*
- *Lexicologie*
- *Linguistique computationnelle*
- *Linguistique de corpus*
- *Littérature*
- *Morphologie*
- *Neurolinguistique*
- *Phonétique*
- *Phonologie*
- *Pragmatique*
- *Psycholinguistique*
- *Sémantique*
- *Sémiotique*
- *Sociolinguistique*
- *Syntaxe*
- *Technologies de l'information*
- *Terminologie*
- *Traductologie*
- *Traitement automatique du langage*
- *Troubles langagiers*

Le comité organisateur décidera du format de présentation (communication orale ou affichée) en fonction des papiers retenus. Cependant, toutes les soumissions acceptées seront publiées de la même manière dans les actes de la conférence. En outre, des prix seront attribués à la meilleure présentation orale et au meilleur poster.

Ces JéTou 2017 proposent ainsi une thématique actuelle destinée à ouvrir de nouvelles perspectives de recherche et de collaboration interdisciplinaires. Ces journées seront alors l'occasion pour tous ceux qui le souhaitent de s'interroger, de débattre, et de confronter leurs travaux et leurs réflexions.

Call for Papers

6th Jétou (Journées d'études Toulousaines)

Young Researchers Conference

Université Toulouse Jean Jaurès – campus Le Mirail

Toulouse, France (May 18th & 19th 2017)

The Jétou (Journées d'études toulousaines) is an international symposium aiming at gathering Master and doctoral students and young researchers (who have defended their dissertation within the past three years) together, from the different disciplines of Linguistics, on an open and multidisciplinary theme. This 6th edition is organized by doctoral students from two laboratories in Toulouse University, France:

- CLLE-ERSS (Équipe de Recherche en Syntaxe et Sémantique)
- Octogone-Lordat (Laboratoire de neuropsycholinguistique)

This **6th edition of the Jétou** will be devoted to a reflection on the following theme: Interfaces in Linguistics.

Linguistics is becoming an increasingly interdisciplinary field of study. This is not surprising, given the nature of language itself: different levels of linguistic structure are in constant interaction, and language also interfaces with numerous other fields of human activity. The goal of the JéTou 2017 conference is to bring together young researchers working on **different scientific fields in and around** linguistics. We therefore welcome all submissions that integrate two or more scientific subfields that address relevant linguistic issues. These works can explore the interfaces between different levels of theoretical linguistic description (phonetics, phonology, morphology, syntax, semantics, discourse), the interaction of theoretical linguistics with other language–related fields (language acquisition, language learning, language teaching, translation studies, semiology, etc.), or the interfaces between language studies in general and other scientific fields such as psychology, neuroscience, sociology, anthropology, computer science, etc.

The list of possible domains includes, but is not limited to:

- **Phonetics**
- **Phonology**
- **Morphology**
- **Syntax**
- **Semantics**
- **Semiotics**
- **Sociolinguistics**
- **Pragmatics**
- **Discourse**
- **Lexicology**
- **Lexicography**
- **Terminology**
- **Language acquisition**
- **Language learning**
- **Language teaching**
- **Translation studies**
- **Literature**
- **Cognition**
- **Psycholinguistics**
- **Neurolinguistics**
- **Speech disorders**
- **Natural Language Processing**
- **Computational linguistics**
- **Corpus linguistics**
- **Information Technologies**

The main focus of the conference will be on the works at the interface of: **1) Linguistics and computer science, 2) Linguistics, psychology and neuroscience and 3) Linguistics, language acquisition, language learning and language teaching**; but any work that is interdisciplinary in nature will be given full consideration. Proposals combining theoretical considerations with work on linguistic data are particularly welcome.

Remerciements / *Acknowledgements*

Nous tenons à remercier très chaleureusement Audrey Bürki et Marie Lallier d'avoir accepté l'invitation que nous leur avons adressée. Nous remercions également l'ensemble des membres du Comité Scientifique pour leurs précieuses et attentives relectures de toutes les propositions soumises. Enfin, nous remercions les modérateurs et les memembres du jury, ainsi que toutes les personnes qui ont apporté leur contribution à l'organisation de ces journées.

We wish to express our deepest gratitude to Audrey Bürki and Marie Lallier for accepting our invitation and being our two guest speakers. We thank all the members of the Scientific Committee for carefully reading and correcting all the submissions. We also acknowledge the moderators and jury members, as all those who have helped and contributed to the organisation of this event.

Un merci particulier à nos partenaires : / *Special thanks to our sponsors :*

- Université Toulouse – Jean Jaurès ;
- Département de Sciences du Langage de l'Université Toulouse – Jean Jaurès ;
- Département d'Etudes du Français Langue Etrangère de l'Université Toulouse – Jean Jaurès ;
- Laboratoire CLLE-ERSS (UMR 5263) ;
- Ecole doctorale CLESCO ;
- Unité de Recherche Interdisciplinaire Octogone-Lordat (EA4156) ;

Comité scientifique / *Scientific Committee*

Charlotte Alazar, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Basilio Calderone, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Anne Condamines, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Cecile Fabre, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Bruno Gaume, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Hélène Giraudo, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Cecilia Gunnarsson, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Nabil Hathout, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Lydia-Mai Ho-Dac, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Mélanie Jucla, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Mouna Kamel, Université Toulouse 3- Paul Sabatier, UMR IRIT

Barbara Köpke, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Pierre Largy, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Vanda Marijanovic, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Marie-Paule Péry-Woodley, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Nathalie Rossi-Gensane, Université Lumière Lyon 2, UMR ICAR

Inès Saddour, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Christiane Soum-Favaro, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Dejan Stosic, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Olga Théophanous, Université Toulouse - Jean Jaurès, URI Octogone-Lordat

Juliette Thuiller, Université Toulouse - Jean Jaurès, UMR CLLE-ERSS

Marianne Vergez Couret, School of Modern Languages, Queen’s University, Belfast

Comité d'organisation / *Organisation Committee*

AHUMADA Lyanne
lahumada@univ-tlse2.fr
OCTOGONE-Lordat

BONNEMAISON Karine
karine.bonnemaison@univ-tlse2.fr
CLLE-ERSS

MERDY Emilie
emilie.merdy@univ-tlse2.fr
CLLE-ERSS/Prometil

MILETIC Aleksandra
aleksandra.miletic@univ-tlse2.fr
CLLE-ERSS

MYTARA Kleopatra
kleopatra.mytara@univ-tlse2.fr
OCTOGONE-Lordat

ORIHUELA Karla
karla.orihuela@univ-tlse2.fr
CLLE-ERSS

SOLIER CLARA
clara.solier@univ-tlse2.fr
OCTOGONE-Lordat

TE RIETMOLEN NOÉMIE
Noemie.terietmolen@univ-tlse2.fr
OCTOGONE-Lord

Langue, locuteur et analogie dans l'acquisition-apprentissage linguistique

Redouane BOUGCHICHE
Université Paris-Sorbonne (Paris 4)
Laboratoire *Sens, Texte, Informatique et Histroire* (*Stih*)
redouane.bougchiche@yahoo.fr

Résumé

Apprendre une langue implique la mise en œuvre d'opérations cognitives nécessaires à la perception et l'intégration des savoirs, ainsi qu'au réinvestissement de ces derniers en savoir-faire linguistiques. L'analogie est l'un des processus cognitifs essentiels à la cognition humaine, particulièrement au processus d'acquisition-apprentissage linguistique. A travers les différents appariements formels et/ou structurels entre les acquis et les nouveaux savoirs linguistiques, le locuteur comprend les discours d'autrui et construit ses propres discours en réinvestissant les connaissances acquises à tous les niveaux de l'analyse linguistique. A travers le processus analogique, cette contribution vise à comprendre le fonctionnement linguistique des locuteurs apprenant une langue maternelle ou étrangère.

Mots-clés: analogie, apprentissage, locuteur, raisonnement analogique.

1 Introduction

L'homme apprend naturellement dès sa naissance, en commençant par acquérir la langue. Cette activité est diverse et complexe notamment à cause des processus engagés à cet effet. L'hétérogénéité développementale réside dans la cognition humaine, dans la manière dont les connaissances sont intégrées, mémorisées et surtout réinvesties. Apprendre n'est pas une simple « copie » ou reprise de contenus dans le cerveau, mais cela implique une *construction* et une *consolidation* du savoir et des modèles mentaux pour comprendre, retenir et fixer les informations durablement, si ce n'est à vie. En se basant sur deux paramètres, l'objet d'apprentissage et le savoir préalable, le sujet distinguera facilement les éléments importants à mettre en relation, synthétisera et structurera les nouvelles connaissances qu'il faut intégrer à celles antérieurement acquises.

En psychologie cognitive, on insiste sur le rôle prépondérant des acquis antérieurs pour les apprentissages futurs, et sur le fait que les nouvelles connaissances sont construites à partir des connaissances existantes. Dans cette conception, le sujet interprète les nouvelles informations en fonction de ce qu'il connaît, et si *« les conceptions initiales vont dans le même sens que les informations nouvelles, l'acquisition sera facilitée »* [1]. Pour comprendre le développement linguistique, il faut étudier les processus cognitifs chez le locuteur : l'analogie est l'un de ces mécanismes.

2 Analogie linguistique et raisonnement analogique

L'analogie[2], dans sa forme originale, désigne l'égalité des rapports de grandeurs mesurables. Elle a d'abord été nommée *proportion* par Pythagore. Suivant la tradition aristotélicienne, l'analogie est constituée de quatre termes *A*, *B*, *C* et *D*, schématiquement, $A : B = C : D$. Autrement dit, la relation qui existe entre les termes *A* et *B* est similaire à celle entre *C* et *D*. Cette catégorie d'analogie est intéressante à étudier dans le cadre de la production linguistique, car elle permet le calcul d'une inconnue jamais produite par le locuteur à travers ce qu'il connaît de la langue. Depuis l'antiquité, l'analogie a été au centre des

[1] LABRELL, F. & MEGALAKAKI, O., 2008, p. 1.
[2] Il existe plusieurs catégories d'analogies linguistiques (voir Monneret, 2004).

discussions grammaticales, qu'elle s'attache à la conjugaison des verbes, à la nature des mots ou aux déclinaisons.

L'analogie joue un rôle dans la régularisation des formes irrégulières. En ancien français, le verbe *trouver* prenait à la première personne du singulier la forme *je treuve*, alors qu'à la première personne du pluriel, il se présentait comme *nous trouvons*. La forme du singulier a fini par se conformer à celle du pluriel pour devenir *je trouve*. La même procédure s'est réalisée pour le verbe *prouver* comme le montre Saussure (1967, p. 222) : *« en français, on a dit longtemps : il preuve, nous prouvons, ils preuvent. Aujourd'hui on dit il prouve, ils prouvent, formes qui ne peuvent s'expliquer phonétiquement ».*

En morphologie verbale par exemple, on distingue les rapports suivants :

• L'analogie dont la similarité est *duelle* telle que : *marcher* : *je marche* = *chercher* : *je cherche*, de même que *marcher* : *chercher* = *je marche* : *je cherche* ;

• La similarité est *simple* comme dans *il marche* : *il marcha* = *il voit* : *il vit* ;

• et dont la similarité entre les éléments est *nulle* telle que *il va* : *il alla* = *il voit* : *il vit*.

En lexico-sémantique, on observe les rapports analogiques dans ce qui suit :

• Un des aspects de cette analogie consiste dans la construction d'éléments qui ne relèvent pas du même paradigme de dérivation affixale. La relation entre *champignon* et *fongique* est la même que celle entre *relation* et *relationnel*. C'est bien le rapport lexico-sémantique qui est mis en avant par cette analogie, où *fongique* sert d'adjectif qualificatif pour *champignon*, tout comme *relationnel* l'est pour *relation*.

• Là où les éléments constitutifs appartiennent au même couple de famille morphologique, autrement dit, avec un rapport formel exprimé entre le moyen et son utilisateur : *piano* : *pianiste* = *violon* : *violoniste*.

• Le rapport relationnel concerne le sens qui lie les lexèmes tels que : *texte* : *écrire* = *maison* : *bâtir*. On distingue la relation sémantique entre le verbe et son action sans aucun rapport formel.

• L'analogie lexico-sémantique a une portée plus large qu'en morphologie, par exemple. Il n'y a pas nécessairement besoin de plusieurs points communs entre les éléments de l'analogie pour qu'elle se réalise comme dans *eau* : (lit de) *rivière* = *circulation* : *rue*. Dans ce cas, c'est la relation instaurée par *s'écouler* ou *contenir* qui est mise en avant car, *l'eau s'écoule le long d'une rivière* ; *la circulation s'écoule le long d'une rue*, et que *la rue contient la circulation*, de même que *la rivière contient l'eau*.

L'analogie participe pleinement à la production de nouvelles phrases. D'après Bloomfield, la construction de phrases relève d'une opération de substitution. Il suffit, pour un locuteur, d'avoir rencontré une phrase à laquelle il substitue d'autres éléments pour avoir plusieurs autres phrases. Celles-ci sont conçues sur le même modèle sans jamais les avoir entendues auparavant. Pour cet auteur : *« les analogies régulières d'une langue sont des habitudes de substitution. Supposons, par exemple, qu'un locuteur n'ait jamais entendu la forme Donne l'orange à Annie mais qu'il ait entendu ou prononcé une série de formes comme celles qui suivent :*

Bébé a faim. Pauvre Bébé ! L'orange de Bébé. Donne l'orange au bébé !

Papa a faim. Pauvre Papa ! L'orange de Papa. Donne l'orange à Papa !

Bill a faim. Pauvre Bill ! L'orange de Bill. Donne l'orange à Bill !

Annie a faim. Pauvre Annie ! L'orange d'Annie……

Il a l'habitude maintenant - l'analogie - d'utiliser Annie dans les mêmes positions que Bébé, Papa, Bill et par conséquent, dans la situation qui convient, énoncera la forme nouvelle Donne l'orange à Annie ! »[3].

Les formations par *effet de couple* (Marchello-Nizia, 2006, p. 85) de syntagmes constituent une autre manifestation de l'analogie syntaxique. En français, *avant que* est fait sur le modèle de *après que*, et fonctionne avec le mode subjonctif. C'est également à travers l'analogie que *je m'en rappelle* a été créé sur le modèle de *je m'en souviens*.

Ainsi, pour être productif, soit le locuteur se base sur des règles toutes faites, soit il mobilise des connaissances déjà mémorisées présentes à son esprit. C'est dans le cadre du second modèle que l'analogie est intéressante à étudier car, d'une part, le locuteur ignore les normes grammaticales de la langue qu'il apprend, et d'autre part, c'est en s'appuyant sur les situations d'apprentissage spontané non-guidé que le langage humain a été construit, et qu'en l'absence des institutions scolaires, les locuteurs apprennent leur(s) langue(s) par transmission en situation, et non par règles conscientes. Envisager l'analogie dans l'apprentissage-production linguistique, c'est l'envisager d'un point de vue cognitif, en tant que processus. En effet, la production analogique est d'ordre psychologique et

[3] BLOOMFIELD, L., 1970, p. 258.

grammatical, *elle suppose la conscience et la compréhension d'un rapport unissant les formes entre elles* (Saussure, 1967 : 226).

L'analogie permet de résoudre des problèmes linguistiques, et acquérir de nouvelles compétences, en recourant à une compétence similaire. Le raisonnement analogique permet de trouver une similitude entre deux situations, de découvrir le lien, structurel/relationnel, existant entre des savoirs acquis : entre *A* et *B*, de manière à ce qu'il permette de former le même rapport entre deux autres éléments : *C* et *D*. Le processus analogie est important pour le fonctionnement cognitif humain, notamment à travers la résolution de problèmes (Gentner 1983, 1989). L'appariement (*mapping* pour Gentner) permet de rapprocher les termes d'une analogie, et leur mise en correspondance permet de résoudre le problème posé. Ce raisonnement permet le calcul d'une inconnue, ce qui est à l'origine de la création. L'analogie est également un des processus intervenants dans la construction du langage (Tomasello, 2003). A travers elle, on comprend comment le locuteur crée et produit dans la langue en fonction de ce qu'il a entendu dans sa communauté linguistique. Le locuteur développe sa capacité à imiter les locuteurs experts, non seulement dans la forme du discours, mais aussi dans l'intention de communication (Tomasello, 2003). Ainsi, le locuteur produit ses propres discours en construisant des schèmes servant de base analogique à de nouvelles constructions.

2.1 L'analogie processus de production linguistique

La productivité langagière s'appuie sur les expériences linguistiques antérieures des locuteurs qui permettent une production automatisée de modèles tout faits (exemplaires) adaptés aux différentes situations de communication. Ces exemplaires servent à construire de nouvelles productions jamais réalisées, autrement dit, des connaissances en construction. C'est dans le processus analogique que se trouve la clef du fonctionnement du locuteur que ce soit dans un cadre monolingue, ou dans un cadre bilingue. Le locuteur apprenant le français imite les formes et les structures linguistiques fournies par ses interlocuteurs (Tomasello, 2003). Le locuteur rapproche les données linguistiques qu'il entend de ses primo-savoirs afin de les comprendre. Dans le cadre de nouvelles productions, il s'appuie sur ces exemplaires afin de trouver le modèle adéquat pour une production personnelle. Le résultat de cette démarche représente une création personnelle.

L'interaction avec des locuteurs, experts ou novices, permet à l'apprenant d'imiter les formes et les structures linguistiques fournies par ses interlocuteurs (Tomasello, 2003). Dans cette situation, le locuteur rapproche les nouvelles données linguistiques qu'il entend de ses primo-savoirs afin de les comprendre. C'est durant cette période qu'il construit des schèmes mentaux représentationnels. Puis, dans le cadre de nouvelles productions, il s'appuie sur ces exemplaires afin de trouver le modèle adéquat pour une production personnelle. Dans cette démarche, il s'agit, pour le locuteur, d'une création. Enfin, si cette dernière n'est pas rejetée par ses interlocuteurs, elle finira par intégrer les paradigmes ou les réseaux de savoirs préconstruits. Ainsi, dans une perspective analogique, le locuteur s'appuie sur des exemplaires rencontrés et mémorisés pour comprendre et produire de nouveaux énoncés (Lavie, 2003 ; Tomasello, 2003).

Comme l'avance Lavie (2003 : 9), « *la productivité est donc la possibilité de produire ou comprendre une infinité d'énoncés dans un cadre linguistique donné, c'est-à-dire à "compétence" constante* ». Dans le cadre de l'appropriation linguistique, on peut comprendre que le locuteur procède à des productions structurelles, car il accède aux savoirs linguistiques par rapprochement des différents éléments formels qu'il reçoit. Puis, avec le développement de sa compétence linguistique, il accède au stade de productivité systémique où il ne se focalise pas seulement sur les ressemblances morphologiques, syntaxiques, etc. pour produire dans la langue, mais il opère des ponts entre les savoirs acquis pour construire des analogies sans ressemblances formelles, telles que les analogies cognitives et lexico-sémantiques *ta mère* : *toi* = *ma mère* : *moi* ; *il va* : *il alla* = *il voit* : *il vit* ; *champignon* : *fongique* = *relation* : *relationnel* ; *texte* : *écrire* = *maison* : *bâtir*, etc.

Le locuteur développe ainsi une nouvelle compétence qui lui permet d'accéder à un autre niveau de production dans la langue sans pour autant se focaliser uniquement sur les similarités formelles. Ainsi, il passe de *marcher* : *je marche* = *manger* : *je mange* à *il est* : *je suis* = *il va* : *je vais* ou *j'irai* : *je vais* = *je mangerai* : *je mange*, entre autres. C'est à travers l'analogie systémique que les dernières constructions sont possibles. Dans le cadre d'une productivité structurelle, le modèle suivi en morphologie, par exemple, est le suivant : *base verbale* + *flexion* =

forme verbale fléchie. Si cette forme convient à beaucoup de constructions analogiques, elle présente certaines limites quand il s'agit des verbes à base verbale différente (allomorphes), selon le temps exprimé par exemple : *irai* est à *vais* comme *mangerai* est à *mange*. Cet exemple se répète avec les verbes *être* et *aller* tels que : *êtes* est à *suis* ce que *allez* est à *vais*. Il en est de même pour les constructions lexico-sémantiques telles que : *écrire* est à *texte* comme *bâtir* est à *maison*. Dans ces exemples, le locuteur met en relation des éléments de la langue qui n'ont pas la même forme. Il connecte entre eux des mots différents dont la relation est basée sur le sens ou sur une racine verbale différente. Parce que le locuteur ne connaît pas la composition de la langue en sous-catégories, en verbes à trois groupes différents, et en champs lexicaux variés, il fournit un effort cognitif.

Ainsi, la productivité linguistique *« résulte du jeu combiné de la productivité structurelle et de la productivité systémique. […] La productivité structurelle couvre la morphologie et la syntaxe en continuité»* (Lavie, 2003 : 103), la productivité systémique est relationnelle, elle se base sur la relation qui lie les unités linguistiques entre-elles. Ainsi, l'appropriation linguistique est incrémentale. Le locuteur apprend la langue par paliers et les savoirs par paradigmes, puis accède au niveau systémique où il associe les savoirs deux par deux de sorte à relier les unités linguistiques entre-elles sans rapports formels, mais que les différentes compositions verbales, lexico-sémantiques et cognitives permettent. Ce dont le locuteur a besoin dans sa pratique linguistique, c'est d'arriver à exprimer dans la langue ce qui est possible et ce qui ne l'est pas (Lavie, 2003 : 17).

2.2 Analogie et apprentissage linguistique

De ce qui précède, nous avons vu quelques aspects de la production linguistique en morphologie et en syntaxe. Cette partie concernera particulièrement le volet lexico-sémantique de l'acquisition-apprentissage des langues et le rôle de l'analogie dans l'usage et la construction du sens linguistique.

2.2.1 Dans un cadre monolingue (chez l'enfant)

Pour communiquer avec autrui, notamment avec l'adulte, le très jeune enfant tente de reproduire ce qu'il entend de l'adulte. Or, jusqu'à deux ans, son développement linguistique ne lui permet pas de construire son discours à l'instar de l'adulte. Il produit des unités limitées du discours d'autrui. Il commence par des productions holophrastiques consistant à produire un mot fonctionnant comme un énoncé entier (Tomasello, 2000), par exemple *balle* pour *je veux, donne-moi la balle*. Ensuite vient la période des multi-mots où l'enfant produit des énoncés tels que *Où est X ? Je veux Y*, etc. L'enfant s'appuie sur ces modèles, également appelés schèmes cognitifs (Tomasello, 2003, Bougchiche, 2013), pour produire d'autres énoncés lui permettant de satisfaire ses besoins expressifs. C'est une étape de productions syntaxiques analogiques jamais réalisées par le passé telles que *je veux une balle*, *je veux une pomme*, *où est maman*, *où est doudou*, etc.

Le développement linguistique mène l'enfant à utiliser les savoirs acquis pour transmettre du sens. La composante sémantique joue un rôle fondamental dans la recherche de synonymes. Dans une situation d'ignorance linguistique, l'enfant cherche dans son lexique mémorisé les correspondances nécessaires pour couvrir ses besoins linguistiques. L'aspect perceptuel des entités (ou réalités) se fait par la recherche des équivalences, dans le cadre d'une analogie binaire, entre deux unités, entre *grand* et *long* ; ou *vase* (A) et *bol* (B) par exemple. La relation entre ces deux derniers termes réside dans le fait que l'un et l'autre sont faits pour contenir un liquide. Ce même phénomène s'observe dans l'utilisation des verbes. Dans « je déshabille la pomme », consistant à *ôter* quelque chose que le verbe *déshabiller* (A) partage avec *éplucher* (B), l'enfant étend le sens de l'un à l'autre du fait que l'action des deux verbes est similaire. L'appariement analogique est fait entre les traits sémantiques partagés par les deux verbes, car l'enfant, ignorant le lexème *éplucher*, se représente l'action d'*ôter* une couche enveloppante, de la même manière qu'il se déshabille lorsqu'il ôte ses habits.

L'enfant trouve dans la synonymie la possibilité combinatoire dans la transmission du sens. Dans **je te parle quelque chose*, l'enfant transpose les possibilités combinatoires de *dire* sur *parler* qui ont, par ailleurs, les mêmes traits sémiques, alors que dans d'autres couples de mots, une seule ressemblance sémantique peut servir à faire d'un mot un usage synonymique : *entendre* et *écouter*, etc. Dans cette action, l'enfant se focalise sur les sèmes communs des mots et leur ressemblance combinatoire, sans prendre connaissance des différents traits séman-

tiques qui les distinguent[4]. Le recours aux traits sémantiques communs se réduit au fur et à mesure que l'enfant accède aux subtilités de la langue, les usages deviennent ainsi de plus en plus spécifiques. La synonymie aura une nouvelle fonction, celle de paraphraser un énoncé, ou de désigner uniquement les entités qui partagent les mêmes traits sémantiques. Progressivement, l'enfant abandonne les mauvais choix lexicaux pour restreindre leurs usages et les modifier, comme nous le montre Oléron : *« la correction des extensions résulte de l'adjonction de traits nouveaux. Grâce à cette adjonction, le mot ne va plus désigner que les objets qui manifestent le nouvel ensemble de traits (chat sera réservé pour chat et tigre, et chien pour chien et loup par exemple) »*[5].

Par ailleurs, en utilisant des verbes réfléchis, l'enfant crée de nouvelles formes pour des verbes qui n'en ont pas besoin, par exemple **tu vas te mourir* pour *tu vas te tue* (Grégoire, 1947 : 171) L'enfant transpose les usages du verbe *tuer*, *je vais me tuer*, *tu vas te tuer*, *il va se tuer*, etc., à ceux du verbe *mourir* dans *je vais mourir*, *tu vas mourir*, *il va mourir*[6] par analogie synonymique. Il aligne les formes du verbe *mourir* sur celles de *se tuer* pour obtenir la forme **tu vas te mourir*. Par cette action, il a introduit la forme réfléchie à ce verbe qui partage des traits sémiques avec son analogon pour signifier la même chose, se donner la mort. L'acquisition sémantique permet d'accéder à la construction abstraite de la langue.

A travers la synonymie, l'analogie sémantique entre lexèmes facilite leur utilisation et leur acquisition. Ainsi, l'apprentissage lexical permet le développement d'abstractions, et la similitude de l'*input* peut être reconnue sur la base formelle des mots en cours d'acquisition. L'analogie permet de dégager une similitude sémantique, notamment par la nature référentielle du lexique où, en dehors des noms propres, tout mot se prête à assumer une fonction générique, comme le montre Oléron : *« en dehors des noms propres, tout mot d'une langue a un caractère générique : il s'applique à des référents qui ne sont jamais identiques (et qui même s'ils l'étaient n'en seraient pas moins multiples). Les normes linguistiques définissent - non sans marges de variation et d'incertitude - le champ des référents auquel chaque mot doit s'appliquer. Il y a extension quand le locuteur étend ce champ et sous-extension quand il le restreint - faisant entrer dans le champ plus d'objets qu'il n'est admis dans le premier cas et moins dans le second »*[7].

L'apprentissage par traits sémantiques (Clark, 1973a/b), est un des modèles théoriques de l'apprentissage sémantique. Malgré plusieurs critiques négatives, ce modèle est encore d'actualité. Il représente une voie explicative du rôle de l'analogie dans l'apprentissage lexico-sémantique. En effet, l'enfant apprend les mots avec une partie de leurs traits sémantiques. Quand il entend *sauter du plongeoir* (ex. : il saute du plongeoir), il se représente l'action de « plonger dans une piscine ». Puis, il entend *sauter un repas* (ex. : il saute un repas), il ajoute la nouvelle acception « ne pas manger », à celle de *plonger*. Il observe que le verbe *sauter* représente un sens différent dans les deux énoncés, il comprend que le mot est polysémique. Par cette action, il envisage d'autres mots, verbes et/ou noms, avec des usages polysémiques. Il comprend que certains mots sont polysémiques. Par exemple, *jumelles* renvoie à la fois à *l'instrument d'optique portatif qui permet de voir de loin et de rapprocher des objets* et à *deux sœurs nées le même jour d'une même maman*. La même représentation sémantique sera observée lorsque le locuteur acquiert les différents sens de *feuille* dans *les feuilles tombent de l'arbre* et *je dessine une maison sur une feuille*.

Ces rapprochements analogiques par traits sémantiques aident le locuteur à choisir un lexème disponible pour celui qu'il ignore sur la base des acquis et usages maîtrisés. Le locuteur s'appuie ainsi sur la similarité partagée par une paire lexicale (ex.: *savoir*/*connaître*), dont il ne perçoit pas la différence au niveau de la combinatoire sémantique. Il ne reconnaît qu'une caractéristique sémantique et combinatoire commune aux deux verbes, ce qui atteste le transfert de l'usage (les usages) de l'un à celui (ceux) de l'autre. Jusqu'à neuf ans, *« l'enfant connaît certains contextes appropriés à la production du mot, mais il n'a pas encore isolé les traits sémantiques »*, c'est de cette façon que l'acquisition sémantique peut être considérée *« comme une abstraction progressive des éléments de signification »* (Bernicot, 1981 : 23). L'enfant conçoit les mots avec leur sens général, puis il parvient à les distinguer en rajoutant des traits sémiques spécifiques pour

[4] Ce qui s'explique également par le fait que l'enfant ignore ce qui distingue ces mots.

[5] OLÉRON, P., 1979, p. 119.

[6] Le français offre la possibilité d'utiliser la forme *se mourir* « être en train de mourir » (*Tlfi*, article *Mourir*). Or, à cet âge, l'enfant ne maîtrise pas cette forme. Ainsi, il a agi par analogie pour la produire, et que le sens de *se tuer* et *se mourir* n'est pas identique.

[7] OLÉRON, P., 1979, p. 85.

chaque usage (Clark, 1973a/b). Pour Clark, l'enfant apprend la signification de certains verbes en ajoutant progressivement des traits sémantiques les uns aux autres. Dans un premier temps, il attribue à tous les mots (dont les verbes sus-cités) un trait général, puis, dans un second temps, il augmente les traits composant ces mots pour devenir plus spécifiques (Bernicot, 1981). En accédant aux usages spécifiques des mots, l'enfant cessera de les confondre et abandonnera les extensions inappropriées pour utiliser le lexème adéquatement. Ce sont des extensions analogiques temporaires qui se dissipent une fois que l'usage approprié acquis.

Plus un locuteur parvient à attribuer un sens (précis) à un mot, plus il sera facile de l'apprendre. Le degré de ressemblance d'un lexème avec un autre mot offre les possibilités de poser des correspondances par analogie entre les mots et de montrer l'influence de la forme sur le sens. À ses débuts, l'enfant n'accorde qu'une signification réduite aux mots par rapport aux significations des usages lexico-sémantiques utilisés par les adultes, car : *« la signification des mots n'est pas donnée d'emblée à l'enfant. Au début du développement, dans bien des cas la signification attribuée à un mot par un enfant ne correspond que partiellement à la signification adulte »*[8]. À force de procéder à des analogies sémantiques, l'enfant applique ce procédé à d'autres mots qu'il acquiert, et il *« acquiert les mots qui correspondent à des référents dans le monde. L'un des premiers apprentissages de l'enfant est celui des concepts catégoriels : citer le mot « chien » réfère non seulement au chien de la maison, mais aussi à tous les animaux de la classe de chien »*[9].

L'enfant crée des appariements analogiques entre le nom et l'objet qu'il représente. Dans cette action, l'enfant peut attribuer le nom *chien* à un *lapin*, au *chien du voisin*, etc. Comment peut-on expliquer ce type de catégorisation sémantique ? L'enfant commence par acquérir un mot, *chien* (A) par exemple, lié à un référent individuel *le chien de la maison* ou *le chien de la voisine*, puis il associe cette représentation au *chien de la télé* ou *au chien de l'image qui se trouve sur le livre que lui lit sa maman* (B). Il peut ainsi associer certains sèmes connus *chien* par leurs ressemblances sélectives de traits physiques communs avec d'autres animaux, les poils, le museau, les pattes par exemple. Il finit par attribuer le nom d'un animal à plusieurs animaux : *chien* (A) pour *lapin* ou *chat* (B), etc.

Il peut également procéder à des restrictions sémantiques[10] en donnant un nom générique à un objet qui renvoie à plusieurs réalités, mais qui, dans un emploi, est réduit à une seule, par exemple le mot *voiture* qui sera exclusivement employé pour la *voiture familiale*. Il généralise un objet et son appellation à d'autres objets qui partagent quelques traits référentiels, à savoir la catégorie animale. Ces deux derniers points se réduisent progressivement avec l'acquisition de nouveaux mots. Plus l'enfant dispose de vocabulaire, moins il fait appel aux extensions et aux restrictions sémantiques. Les corrections faites par les adultes feront que l'enfant apprend l'appellation correcte et modifie son comportement linguistique. Ainsi, il élargit ses paradigmes et diversifie l'usage de la langue.

2.2.2 Dans un cadre bilingue (chez l'adulte)

Le premier accès à la signification dans la nouvelle langue se fait sur la base des représentations sémantiques de la langue maternelle (désormais L1). Le locuteur se reporte constamment à ces représentations pour construire de nouvelles significations, car *« les significations en langue seconde sont médiatisées par les concepts quotidiens représentés par la langue maternelle »* (Bange, 2005 : 73). En effet, chaque fois que le locuteur veut exprimer une idée dans la nouvelle langue (désormais NL), il se projette dans le système conceptuel de L1, ce qui, parfois, aboutit à des confusions. Le choix lexico-sémantique en NL consiste à généraliser le sens du mot connu en L1 (avec toutes les acceptions connues), à celui de la langue en cours d'apprentissage. Le kabylophone utilise le verbe *manger* en français avec le sens qu'il a en kabyle : « avaler une nourriture après l'avoir mâchée » après l'avoir entendu dans des énoncés : *j'ai mangé une pomme*, *je mange à midi*, *nous mangeons à la cantine*, etc. Il attribuera à ce verbe d'autres acceptions que le français n'autorise pas, à l'instar des expressions idiomatiques de L1. L'exemple suivant est produit par une locutrice kabylophone résidant à Paris : *cigh argaziw* [tʃiɣ/argaziw], litt. « j'ai mangé mon mari », sém. « j'ai enterré mon mari », « mon mari est décédé avant moi ».

Dans cet exemple, l'analogie interlinguistique est due au succès de l'utilisation de la première acception du verbe *manger* avec le sens

[8] BERNICOT, J. & BERT-ERBOUL, A., 2009, p. 57.
[9] *Ibidem.*, p. 57-58.
[10] La restriction sémantique consiste à utiliser un hyperonyme à la place d'un hyponyme (*boisson* pour *alcool* par exemple).

de « avaler », la locutrice s'appuie sur ce premier usage pour en produire le second.

Par ailleurs, l'erreur proviendrait des usages sémantiques des mots de L1 qui influent sur le choix lexico-sémantique de NL. Si L1 accorde le même signifiant pour deux signifiés, cet usage sera transposé à NL. *Réaliser* en français est utilisé à la fois comme : 1. *faire* (quelque chose), et 2. *prendre conscience* (de quelque chose). Le locuteur francophone apprenant l'anglais transposera ces deux usages (1+2) au seul usage (2) correspondant en anglais. Or, malgré la similitude lexicale, l'acception (1) du français ne pourrait être comprise, ou même admise, par un anglophone monolingue.

Le locuteur adulte procède à des extensions, d'une part, semblables à celles de l'enfant (besoin lexical) ; d'autre part, par transfert des acceptions connues en L1. Une traduction littérale des lexèmes de L1 explique généralement ces extensions impropres (voir *réaliser*, *manger*). Les mises en correspondances sémantiques interlinguistiques favorisent ce type d'extensions. La différence de visions du monde dans les langues en contact est une des raisons de l'écart sémantique, car le système conceptuel de la langue en cours d'apprentissage ne peut se résumer à un simple passage (ou une continuité) du système représentationnel de L1. Cependant, le nouveau système ne peut se développer et se parfaire sans que les conceptions de L1 ne soient présentes. Le nouvel apprentissage lexico-sémantique doit répondre aux exigences conceptuelles de NL, dans le cas contraire, il y aura inévitablement écart sémantique. Cette dernière, et à l'instar des autres domaines de l'analyse linguistique, résulte d'abord d'une influence interlinguale (avec une forme ou un sens de L1), ensuite par extension analogique inadéquate du lexique appris en NL.

La reprise d'une forme lexicale connue en L1 en apprenant une NL est une source d'analogies. Dans le cas où un francophone apprenant l'espagnol rencontre le mot *gato*[11] [gato] « chat », sa première interprétation sera basée sur les similarités phonétiques entre le mot espagnol et celui du français d'où il tire le sens : *gâteau*. La situation inverse est aussi vraie. Un hispanophone apprenant le français rencontrant le mot *gâteau*, il le rapporte à la forme et au sens du mot *gato* de l'espagnol. La similarité phonétique appelle une analogie lexico-sémantique qui mène, la plupart du temps[12], à un usage erroné. La même situation se produit entre le finnois et le français. Le locuteur francophone, apprenant le finnois, reprendra la représentation formelle et sémantique du mot *poule* (oiseau), en rencontrant pour la première fois le mot *pulla* dont le sens, « petit gâteau », est différent de celui du français. C'est la similarité phonétique qui conduit le locuteur francophone à se représenter le sens de ce mot par rapport à son correspondant en français *poule*. Le locuteur opère une mise en correspondance basée sur la représentation lexicale et phonétique du mot en L1 et attribue le sens de *poule* à celui de *pulla* comme suit :

poule (mot fr.) : *poule* (idée fr.) = *pulla* (mot fin.) : **pulla* (idée fr.) ; ou bien :

pulla (mot fin.) : *pulla* (idée fin.) = *poule* (mot fr.) : **poule* (idée fin.).

Les locuteurs ont tendance à assimiler les mots de NL avec ceux de L1, particulièrement quand ces mots présentent des similarités formelles (voir *gato*). Si les deux langues ont des affinités lexicales riches (français-anglais, français-italien, en l'occurrence) - ceux qui gardent la forme de départ, et prennent une nouvelle signification, les reprises lexico-sémantiques par analogie renforcent l'apparition des erreurs.

L'analogie interlinguale permet au locuteur de s'appuyer sur des lexèmes formellement similaires, ou d'attribuer une acception d'un mot de L1 à son équivalent en NL. L'analogie bilingue révèle que le locuteur exploiterait les correspondances entre les deux langues plutôt que de mémoriser un nouveau lexique à côté de celui dont il dispose. Cette opération est cognitivement moins coûteuse, mais conduit souvent à des écarts plutôt qu'à des résultats heureux. Cela dit, l'apprenant produit du sens en adéquation avec le système linguistique. L'analogie sémantique ne se réduit pas aux écarts sémantiques. L'extension représente une étape indispensable avant une maîtrise avancée de la sémantique de NL.

3 Conclusion

A travers le processus analogique, le locuteur comprend et apprend la langue. Il devient autoproducteur de son discours, tout en interagissant avec autrui. L'analogie offre au locuteur les moyens de ses productions, à travers les différents appariements qu'il réalise entre les savoirs maîtrisés et les possibilités de productions. Le

[11] À l'écrit, la même situation se produirait en rencontrant *gato* pour la première fois.

[12] Il existe des mots analogues qui renvoient à la même réalité, et dont le sens est identique dans les deux langues : *adivinar* et *deviner*.

locuteur crée et comprend de nouveaux contenus informatifs, et ce qui rend possible telle ou telle nouvelle production dans la masse des connaissances linguistiques. Il devient autonome dans son rapport, sa cohésion et ses échanges avec son environnement linguistique auquel il appartient et dans lequel il évolue. Il devient l'auteur de son discours, et créateur du sens qu'il veut transmettre. Cela vaut pour l'acquisition de L1 et pour l'apprentissage d'une NL, dans un cadre bilingue, cette fois-ci avec une différence, L1 influe sur les structures de NL.

Au fur et à mesure que le locuteur avance dans son apprentissage, il dispose d'un ensemble de savoirs et de savoir-faire linguistiques qu'il réutilise, en les rappelant, dans des situations analogues basées sur ses expériences personnelles antérieures.

L'analogie est un processus permettant de résoudre des problèmes, mais avant tout d'accéder, à la fois, à la langue et au sens transmis par les interlocuteurs. Pour faire face à des situations de communication inédites, le locuteur s'appuie sur les connaissances maîtrisées, et c'est en se focalisant sur ces acquis que l'analogie prend forme et que le locuteur devient productif. L'analogie permet au locuteur de combler les lacunes linguistiques, car le besoin expressif est plus large que les moyens linguistiques dont il dispose. Ainsi, il n'a pas le sentiment d'être en « contradiction » avec ce qui existe dans la langue, même si la création est malheureuse. Pour lui, ce qu'il « crée » n'est pas une innovation, mais une production « conforme » à ce que la langue lui offre comme possibilités de production.

Références

BANGE, P. (2005), *L'apprentissage d'une langue étrangère: cognition et interaction*, Paris : L'Harmattan.

BANGE, P., CAROL, R. & GRIGGS, P. (2002), « La dimension cognitive dans l'apprentissage des langues étrangères », dans *Revue Française de Linguistique Appliquée*, V. VII, p. 17-29.

BERNICOT, J. (1981), *Le développement des systèmes sémantiques de verbes d'action*, Paris : Editions du CNRS.

BERNICOT, J. & BERT-ERBOUL, A. (2009), *L'acquisition du langage par l'enfant*, Paris : Ed. IN PRESS.

BLOOMFIELD, L. (1970 [1933]), *Langage*, (traduit par Gazio Janick), Paris: Payot.

BOUGCHICHE, R. (2013), *L'analogie dans l'apprentissage des langues*, Thèse de doctorat, Paris4-Sorbonne.

CLARK, E. V. (1973a), « What's in a word? On the child's acquisition in semantics in his first language », in *Cognitive development and the acquisition of language*, New York: Academic Press.

CLARK, E. V. (1973b), « Non-linguistic strategies and the acquisition of word meanings », in *Cognition 2*, 161-182.

GENTNER, D. (1989), «The mechanisms of analogical learning», in *Similarity and analogical reasoning*, New York: Cambridge University Press, p. 197-241.

GRÉGOIRE, A. (1947), *L'apprentissage du langage II*. Bruxelles : Duculot.

HOFSTADTER, D. & SANDER, E. (2013), *Analogie, cœur de la pensée*, Paris : Odile Jacob.

HOLYOAK, K. J. (1985), «The pragmatics of analogical transfer», in G.H. Bower (Ed.), *The psychology of learning and Motivation*, V. 19, New York: New York Academic Press, p. 59-87.

LABRELL, F. & MEGALAKAKI, O. (2008), *Psychologie française*, Issy les Moulineaux : EMSAS.

LAVIE, R.-J. (2003), *Le locuteur analogique ou la grammaire mise à sa place*, Thèse de doctorat, Paris X-Nanterre.

MARCHELLO-NIZIA, Ch. (2006), *Grammaticalisation et changement linguistique*, Bruxelles : De Boeck & Larcier.

MONNERET, Ph. (2004), *Essais de linguistique analogique*, Dijon : A.B.E.L.L.

OLÉRON, P. (1979), *L'enfant et l'acquisition du langage*, Paris : PUF.

SAUSSURE, F. de (1967 [1916]), *Cours de linguistique générale*, Paris : Payot.

TOMASELLO, M. (2000), «First steps toward a usage-based theory of language acquisition», in *Cognitive Linguistics*, Walter de Gruyter, pp. 61-82.

TOMASELLO, M. (2003), *Constructing a Language. A Usage-Based Theory of Language Acquisition,* Boston: Harvard University Press.

First language attrition at two interfaces:

Binding Interpretations of *ziji* 'self' by Chinese-English bilinguals

Wenjia Cai
The University of Edinburgh
Dugald Stewart Building
3 Charles Street Lane
EH8 9AD
wenjiacai09@gmail.com

Abstract

The current study investigates the L1 attrition effects in binding interpretations of *ziji* 'self', among Chinese-English late bilinguals living in the second language environment. The data will be collected from a speeded-online-comprehension task (2AFC), a battery of tests of executive functions (Foster et al., 2015), followed by a sociolinguistic questionnaire (Schmid & Dusseldorp, 2010). According to previous studies of native Chinese speakers, the locality effect was shown during online interpretations of *ziji*. Based on the assumptions that local binding requires less cognitive resources than long-distance binding, and that anaphoric dependencies partially draw on the same pool of attentional resources used to keep two languages separate (Sorace, 2016), I expect that Chinese-English bilinguals with bigger length of residence (LoR) will be more likely to refer *ziji* to a local antecedent, regardless of the discourse context. I also expect the binding interpretations to be influenced by the individual differences in executive functions.

Keywords: L1 syntactic attrition, Interface, Reflexive pronoun, Executive functions, Late bilinguals, Chinese

1 L1 attrition in the pronominal system

Recent studies have shown that extensive exposure to a second language (L2), accompanied by long-term disuse of a first language (L1) could induce restructuring in the syntactic module of the L1 grammar, albeit slowly and selectively (Chamorro, Sorace, & Sturt, 2015a; Chamorro, Sturt, & Sorace, 2015b; Gürel, 2004; Kim, Montrul, & Yoon, 2010; Tsimpli, Sorace, Heycock, & Filiaci, 2004). The selective nature of L2-induced change in the L1 syntactic module has been one of the primary concerns in L1 attritions studies.

The Interface Hypothesis, proposed by Sorace and her colleagues in 2006, is one of the few theories that combines both linguistic and psycholinguistic accounts when explaining the L1 attrition effects. They argued that compared to structures within the core grammar, structures at the interface between syntax and other cognitive domains, for example, the interface of syntax and discourse, syntax and pragmatics, are more vulnerable to language attrition. They also argue that the effects of attrition do not involve the representation of syntactic knowledge, but rather the processing strategies, and the ability to integrate different information in real-time (Sorace, 2011). In fact, one of the reasons that "interface structures" behave differently from others, is that integrating information across different cognitive domains in real-time puts a strain on participants' limited cognitive resources; meanwhile inhibiting irrelevant information from the other language already consumes a lot of resources (Green, 1986), leaving the participants performing at a sub-optimal level.

The prediction made by *Interface Hypothesis* has been supported by a series of studies investigating the bilingual pronominal system (see Sorace, 2011 for a review); among which only the Chamorro studies (2015a, 2015b) and the Tsimpli study in 2004 concerned themselves with the L1 attrition of the late bilinguals, while other studies mainly focused on early bilinguals or heritage speakers. The current study aims to fill this gap by investigating the L1 attrition effects among late

Chinese-English bilinguals, to observe how full-fledged L1 is influenced by L2, without the compound influence of incomplete acquisition.

In addition, the cognitive aspect of the *Interface Hypothesis* hasn't been thoroughly explored as the linguistic aspect: to what extent can we attribute the selectivity of L1 attrition in certain linguistic structures, to the change in cognitive control abilities? By introducing the shortened complex span-test developed by the Engle Lab (Foster et al., 2015), as well as the Test of Everyday Attention (Robertson et al., 1994), the current study hopes to establish a more direct link between the cognitive control abilities and the selectivity in L1 syntactic attrition.

Finally, to control the variations of sociolinguistic factors that may interact with the cognitive control abilities, I will follow the practice of Schmid and Dusseldorp (2010) and closely monitor the pattern of bilingual language use, including but not limited to: LoR, the amount of language use, types of language use, and affiliations to both languages and cultures.

2 Who is *ziji*?

2.1 Chinese reflexives

In Mandarin Chinese, there are two types of reflexives, one is the bare reflexive *ziji* 'self'; the other is the compound reflexive, which combines *ziji* 'self' with a pronoun, e.g. *ta ziji* 'himself', *wo ziji* 'myself', *nimen ziji* 'yourselves'. The compound reflexive behaves in a similar way with its English counterpart. For example (sentences cited from Huang, Li, & Li, 2008):

(1) *Zhangsan$_i$ zhidao Lisi$_j$ lao piping taziji$_{i/*j}$.*
Zhangsan know Lisi incessantly criticize himself
"Zhangsan$_i$ knows that Lisi$_j$ criticizes himself$_{i/*j}$ all the time."
(2) *Zhangsan$_i$ zhidao Lisi$_j$ renwei taziji$_{i/j}$ zui congming.*
Zhangsan know Lisi think himself most clever
"Zhangsan$_i$ knows that Lisi$_j$ thinks he$_{i/j}$ is the smartest."

However, when it comes to the reflexive in its bare form, it's not always bound within its local domain, as suggested by the *Binding Principle* (Chomsky, 1981). While local binding (LOC) is always possible (give that local binder is available), long-distance binding (LD) can appear under certain circumstances, thus causing ambiguity when there are more than one potential antecedents (Huang et al., 2008).

(3) *Zhangsan$_i$ zhidao [Lisi$_j$ chang zai bieren mianqian piping ziji$_{i/j}$].*
Zhangsan know Lisi often at others face criticize self
"Zhangsan$_i$ knows that Lisi$_j$ often criticize him$_{i/*j}$/himself$_{*i/j}$ in the presence of others."
(4) *Zhangsan$_i$ xiangxin [Lisi renwei [ziji$_{i/j}$-de erzi zui congming]].*
Zhangsan believe Lisi think self-DE son most clever
"Zhangsan$_i$ believes Lisi$_j$ thinks that his$_{i/j}$ son is the smartest."

The ambiguity can be resolved, using discourse information that favors either a local or a distant antecedent. For example:

(5a) *Lisi$_j$ xihuan zai beihou yilun bieren, Zhangsan$_i$ zhidao*
Lisi like at behind-the-back judge others, Zhangsan know
"Lisi$_j$ likes to judge others behind their back, Zhangsan$_i$ knows
[Lisi$_j$ chang zai bieren mianqian piping ziji$_{*i/j}$].
Lisi often at others face criticize self.
Lisi$_j$ often criticize him$_{*i}$/?himself$_j$ in the presence of others."
(5b) *Lisi$_j$ hen shanyu fanxing, [Zhangsan$_i$ zhidao Lisi$_j$ chang*
Lisi very good-at self-reflection, Zhangsan know Lisi often
"Lisi$_j$ is very good at self-reflection, Zhangsan$_i$ knows Lisi$_j$ often
zai bieren mianqian piping ziji$_{*i/j}$]
at others face criticize self
criticize ?him$_{*i}$/himself$_j$ in the presence of others."

The current study focuses on reflexive in its bare form *ziji* 'self', which differs from its English counterpart, in the way that it can refer to the distant antecedents beyond the local domain.

2.2 Semantic constraints of long-distance binding

Apart from the discourse information, the semantic meaning of the verb can restrain the long- distance binding of *ziji*. Jin (2003) classified the Chinese transitive verbs into two categories according to whether these verbs can take *ziji* as an object. If the verb in a simple subject-verb-object (SVO) sentence cannot take *ziji* as an object, i.e. if the agent and the patient of the verb cannot be the same person, like in sentence (6), then when this SVO sentence is used as a subordinate clause, like in sentence (7), *ziji* can only be referring to the matrix subject. On the other hand, if the verb in a simple S-V-O sentence can only take *ziji* as an object, i.e. if the agent and the patient of the verb must be the same person, like in sentence (8), then when this SVO sentence is used as a subordinate clause, like in sentence (9), *ziji* can only be referring to the local subject (Jin (2003), cited from Li & Zhou, 2010). The first category is called reflexive verb, while the second is called non-reflexive verb (cited from Li & Zhou, 2010, p. 98). There is a third category in which the verb can take both reflexive and non-reflexive as an object, and in this way, if the SVO sentence is used as a subordinate clause, *ziji* can refer to either the local or the matrix subject,

causing ambiguity; like in sentence (10).

(6) *Zhangsan$_i$ huida ziji$_{i/*j}$.*
Zhangsan answer self
* "Zhangsan$_i$ answered self$_i$."

(7) *Lisi$_j$ rang Zhangsan$_i$ huida ziji$_{*i/j}$.*
Lisi ask Zhangsan answer self
"Lisi$_j$ asked Zhangsan$_i$ to answer him$_{i/*j}$/*himself$_{*i/j}$."

(8) *Zhangsan$_i$ tanbai ziji$_{i/j}$.*
Zhangsan confess self
Zhangsan$_i$ confessed himself$_i$.

(9) *Lisi$_j$ rang Zhangsan$_i$ tanbai ziji$_{*i/j}$.*
Lisi ask Zhangsan confess self
"Lisi$_j$ asked Zhangsan$_i$ to confess *him$_{*i/j}$/himself$_{*i/j}$."

(10) *Lisi$_j$ rang Zhangsan$_i$ buyao shanghai ziji$_{*i/j}$.*
Lisi ask Zhangsan not-to hurt self
"Lisi$_j$ asked Zhangsan$_i$ not to hurt him$_{i/*j}$/himself$_{*i/j}$."

2.3 Locality effects during online processing

Despite that long-distance binding of Chinese reflexive *ziji* is formally possible, many experimental studies have shown that Chinese native speakers displayed local preference when processing *ziji* online. For example, Li and Zhou (2010) conducted a ERP experiment in Mandarin, measuring the electrophysiological response to the anaphor *ziji* in examples like (11a) and (11b).

(11a) *Xiaoli$_i$ rang Xiaozhang$_j$ buyao weizhuang ziji$_{*i/j}$.*
Xiaoli ask Xiaozhang not-to disguise self
"Xiaoli$_i$ asked Xiaozhang$_j$ not to disguise *him$_{*i/j}$/himself$_{*i/j}$."

(11b) *Xiaoli$_i$ rang Xiaozhang$_j$ buyao qianlian ziji$_{i/*j}$.*
Xiaoli ask Xiaozhang not-to embroil self
"Xiaoli$_i$ asked Xiaozhang$_j$ not to embroil him$_{*i/j}$/*himself$_{i/*j}$."

Li and Zhou observed a significantly larger positivity (P300/P600) at *ziji*, when the semantics of the verb blocked the local binding, forcing *ziji* to bind with a distant antecedent, as in (11b); compared to when the semantics of the verb confined *ziji* at its local domain, as in (11a). The results suggested that long-distance binding requires more processing resources (Li & Zhou, 2010).

Cross-model priming studies pointed to a similar advantage for local antecedents over distant antecedents. Gao and colleagues (Gao, Liu, & Huang, 2005; Liu, 2009) presented participants with audio stimuli of the form in (12). Upon reaching the sentence-final *ziji*, participants were presented with a visual probe word. When the probe was presented immediately after the anaphor, participants recognized probes that were semantic associates of local antecedents significantly more quickly; this locality effect disappeared (Gao et al., 2005) or reversed (Liu, 2009) at slightly longer SOAs (160ms or 370ms).

(12) *Zhangsan$_i$ shuo Lisi$_j$ nongshang-le ziji$_{i/j}$.*
Zhangsan say Lisi harm -PFV self.
"Zhangsan$_i$ said that Lisi$_j$ harmed him$_{i/j}$/himself$_{i/j}$."

Using a self-paced reading paradigm, Chen et al. (2012) showed that a locally bound *ziji* was read more quickly than a *ziji* bound with distant antecedent. The results were later replicated in an eye-tracking-while-reading study (Jäger, Engelmann, & Vasishth, 2015).

2.4 Binding interpretations at two interfaces

As we've discussed before, the binding interpretations can be influenced by either discourse or semantic information in the sentence, placing the pronominal structure at the interface of either syntax and discourse, or syntax and semantics. Experimental conditions and exemplar sentences are listed below in Table 1, and a pre-test of the stimuli will be conducted before the experiment, to make sure that the manipulation is consistent with the binding interpretations of native speakers. Details about the experimental materials will be discussed in Section 3.2.

It's worth mentioning that, in the syntax-discourse conditions, binding *ziji* with the less preferable antecedent will not lead to ungrammaticality; it will, however, in the syntax- semantics conditions.

Table 1 Structures of the stimuli

Interface	Interpretation	Sample sentence	Number
Syntax-discourse interface	Local (LOC)	张三$_i$看到李四$_j$站在楼顶，张三$_i$/让李四$_j$/不要/伤害自己$_{i/j}$。 Zhangsan$_i$ saw Lisi$_j$ standing on the roof, Zhangsan$_i$ asked Lisi$_j$ not to hurt himself$_{i/j}$.	N=36
	Long-distance (LD)	李四$_i$拿刀威胁张三$_j$，张三$_j$/让李四$_i$/不要/伤害自己$_{i/j}$。 Lisi$_i$ threatened Zhangsan$_j$ with a knife, Zhangsan$_j$ asked Lisi$_i$ not to hurt him$_{i/j}$.	N=36
Syntax-semantics interface	Local (LOC)	张三$_i$发现李四$_j$在偷东西，张三$_i$/让李四$_j$/坦白自己$_{i/j}$。 Zhangsan$_i$ caught Lisi$_j$ stealing stuff, Zhangsan$_i$ asked Lisi$_j$ to confess himself$_{i/j}$.	N=36
	Long-distance (LD)	张三$_i$一直没收到李四$_j$的回信，张三$_i$/让李四$_j$/回答自己$_{i/j}$。 Zhangsan$_i$ haven't received Lisi$_j$'s reply, Zhangsan$_i$ asked Lisi$_j$ to answer him$_{i/j}$.	N=36

3 Research questions

1) Will the binding interpretations of the short-term group significantly diverge from the long-term group?

2) If the binding interpretation of the Chinese reflexive *ziji* 'self' is sensitive to language attrition, will the pattern of bilingual language use (a multifactor variable consists of LoR, language use and affiliation to both languages and cultures, see Schmid and Dusseldorp, 2010) cast a significant effect on the binding interpretation?

3) If the binding interpretation of the Chinese reflexive *ziji* 'self' is sensitive to language attrition, will the anaphora resolution at the syntax-discourse interface be significantly different from that at the syntax-semantics interface?

4) If the binding interpretation of the Chinese reflexive ziji 'self' is sensitive to language attrition, will the individual differences in executive functions be able to (partially) account for the change in the L1 pronominal system?

4 Experiment procedures

4.1 Participants

As discussed in the first section, data will be collected from both long-term (n=36) and short-term residents in the L2 environment (n=36), to observe the influence of bilingual language experience (LoR, language use, etc.) on one's executive functions (selective attention, and attention switching), which, in turn, can affect the online processing of specific linguistic structures, i.e. reflexive *ziji* 'self'.

Participants in the long-term group are Chinese-English bilinguals who has been living in the L2 environment for more than 7 years, and the short-term group less than 6 months. As late bilinguals, all the participants acquired their second language, and obtained advanced to near- native proficiency (IELT 6.5 or above) after 18 years old.

A sociolinguistic questionnaire adapted from the Schmid and Dusseldorp study (2010) is used to monitor the patterns of bilingual language use, including the amount of language use, types of language use, and affiliations to both languages and cultures. According to Schmid and Dusseldorp, the interplay between the language use and the degree of attrition is far more complicated than previously assumed, and it's the quality rather than the quantity of language use that's crucial to slow down or speed up L1 attrition. They argued that L1 use for professional purpose, which falls into the intermediate mode under Grojean's (1999) model, appears to be the most powerful predictor of L1 attrition, among many other influential factors.

Since the current study anticipates the online processing outcome, as well as the executive function, to be sensitive to the bilingual language experience, we should at least be as cautious to the interplay of all the extra-linguistic factors.

4.2 Materials

A total of 160 sentences were initially created, among which 86 sentences were adapted from the stimuli used in the Li and Zhou study (Li & Zhou, 2010). All the sentences were with the structure of "context sentence + target sentence (P-NP1+VP1+P-NP2+VP2+*ziji*)" (see Table 1). The context sentences were responsible for directing the reflexive reference towards either the matrix subject (P-NP1) or the local subject (P-NP2). Note that in both semantic-biased conditions (Table 1), the semantic meaning of the verb itself were sufficient for making the judgment, and the context sentence merely served the purpose of structural consistency.

10 participants, not-tested for the experiment, will be asked firstly to judge whether *ziji* was referring to the distant reference (P-NP1) or the local reference (P-NP2), and then to rate on a 7-point Likert scale to what extent *ziji* could refer to the distant (point 1) or the local reference (point 7). The sentences chosen for the actual test will be manipulated in such a way that, the local bias will not be qualitatively different in discourse and semantic condition, and so does the distant bias. If there's any difference across the two conditions, it will not be attributed to the artificial effects, but rather the differences in cognitive loads.

In addition, another influential factor "first-mention bias", was considered when designing the stimuli. The first-mention effect was found to be quite effective in pronoun resolution; the study of Chen et al. (2000) showed that in sentences where more than one potential antecedent was present, the antecedent that was mentioned first

will be preferred over the alternative (H. C. Chen, Cheung, Tang, & Wong, 2000). Also, there has been ample evidence that the first- mention effect can be seen beyond the sentence boundaries. Thus, the context sentences were counterbalanced by order of mentioning, that is, half of the context sentence were constructed with matrix subject mentioned first, while half with local subject mentioned first.

4.3 Procedures

During the experiment, participants will be sat in front of a computer in a sound-attenuating experiment booth. Sentences will be presented phrase-by-phrase (Table 1) in the rapid serial visual presentation mode (RSVP) at the center of the screen (see Figure 1). Each phrase consists of 1-2 disyllable words (2-4 characters). Studies have shown that the word-by-word presentation format resulted significantly more accurate reading comprehension than the character-by-character presentation format, probably due to the fact that the Chinese word, rather than the character, is the reading unit (see Lin & Shieh, 2006). All materials will be presented in a white-against black background. Text/background color combination with higher color difference was shown to yield significantly better performances (Wang & Chen, 2003). The presentation rate will be 240 CPM (characters per minute), thus the presentation time for a disyllable word will be 500ms. According to Lin and Shieh (Lin & Shieh, 2006), when the presentation rate was below 240 CPM, a higher level of recall accuracy (nearly 90%) can be obtained. To observe any possible processing deficits, the presentation rate of 240 CPM was chosen to guarantee that participants are reading as fast as they could without compromising reading comprehension.

Figure 1 Experiment procedures

And the end of each sentence, participants will answer whether *ziji* refers to the matrix or local subjects, by pressing the corresponding right/left button on the SR box. The question stays on the screen until the participants made the response or the time limit expires (3000ms). The left/right assignment of response buttons to the binary judgment will be counterbalanced across participants. Sentences were separately by 1000ms interval blank screen.

Sentences will be divided into 4 blocks with 36 sentences in each block. Sentences from each of the four conditions (LD discourse, LOC discourse, LD semantic and LOC semantic conditions) will be distributed equally each block. Sentence in each block will be pseudo-randomized for each participant, with the restriction that sentences with similar verbs or personal pronouns will not be seen in the same block. Participants will be given 3 minutes to rest after each block.

5 Hypothesis and discussion

According to Sorace (2016), bilingual speakers' cognitive abilities in selective attention and/or attention switching are largely enhanced, because they intentionally inhibits irrelevant information from the other language, and constantly switch between two different languages. Based on the assumption that 'anaphora dependencies (partially) draw on the same pool of attentional resources used to keep the two languages separate' (2016, p. 9), there appears to be a "trade-off" between the inhibition abilities and the ability to integrate information from multiple resources in real-time. And this potential trade-off could be the loci of difficulty when processing anaphora dependencies on-line. Thus, the hypothesis of the current study is as follows.

Even though *ziji* in its bare form can be bound outside the local domain, the locality effect during online comprehension indicates that, binding *ziji* with a long-distance antecedent requires more cognitive resources, which should otherwise be used to inhibit irrelevant information: whether it's because of the interference between the local and the distant antecedents, or it's the cross-linguistic influence from the L2 English.

I expect the binding interpretations of reflexive *ziji* will be sensitive to the language attrition; and the group of bilinguals with longer length of residence or those with less L1 use in the intermediate monolingual mode, will be more likely to interpret *ziji* as referring to the local antecedents. However, following the prediction of the *Interface Hypothesis*, because the syntax-

semantic interface is within the formal linguistic modules, participants in either group are unlikely to ungrammatically bind *ziji* with a local antecedent, when the subordinate verb is non-reflexive. Finally, I expect there to be interaction between the cognitive control abilities, the pattern bilingual language use, and the online performance of binding interpretations.

References :

Chamorro, G., Sorace, A., & Sturt, P. (2015). What is the source of L1 attrition? The effect of recent L1 re-exposure on Spanish speakers under L1 attrition. *Bilingualism: Language and Cognition*, 1–13. http://doi.org/10.1017/S1366728915000152

Chamorro, G., Sturt, P., & Sorace, A. (2015). Selectivity in L1 attrition: Differential object marking in Spanish near-native speakers of English. *Journal of Psycholinguistic Research*, *1*, 1689–1699. http://doi.org/10.1017/CBO9781107415324.004

Chen, H. C., Cheung, H., Tang, S. L., & Wong, Y. T. (2000). Effects of antecedent order and semantic context on Chinese pronoun resolution. *Memory & Cognition*, *28*(3), 427–38. http://doi.org/10.3758/Bf03198558

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. *Empirical Approaches to Linguistic Theory: Studies of Seaning and Structure*, 1–20.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Mouton de Gruyter. Retrieved from https://books.google.co.uk/books?id=l08tpkOOdNQC

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–36. http://doi.org/10.3758/s13421-014-0461-7

Gao, L., Liu, Z., & Huang, Y. (2005). Who is ziji: An experimental research on Binding Principle. *Linguistic Sciences*, *4*(2), 39–51.

Green, D. W. (1986). Control, activation, and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language*, *27*(2), 210–223. http://doi.org/10.1016/0093- 934X(86)90016-7

Grosjean, F. (1999). The bilingual's language modes. In J. L. Nicol (Ed.), *One Mind, Two Languages: Bilingual Language Processing.* (pp. 1–22).

Gürel, A. (2004). Selectivity in L2-induced L1 attrition: A psycholinguistic account. *Journal of Neurolinguistics*, *17*(1), 53–78. http://doi.org/10.1016/S0911-6044(03)00054-X

Huang, C.-T. J., Li, A., & Li, Y. (2008). Anaphora. In *The Syntax of Chinese* (1st ed., pp. 329–370). Cambridge: Cambridge University Press. Retrieved from http://www.people.fas.harvard.edu/~ctjhuang/HLL_2007_pdf_folder/HLL2007.html

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*(May), 617. http://doi.org/10.3389/fpsyg.2015.00617

Jin, Z. H. (2003). Verb Restraint Function to ziji long-distance binding. *Chin.Lang.Learn.*, *4*, 9–12. Kim, J.-H., Montrul, S., & Yoon, J. (2010). Dominant language influence in acquisition and attrition of binding: Interpretation of the Korean reflexive caki. *Bilingualism: Language and Cognition*, *13*(1), 73. http://doi.org/10.1017/S136672890999037X

Li, X., & Zhou, X. (2010). Who is ziji? ERP responses to the Chinese reflexive pronoun during sentence comprehension. *Brain Research*, *1331*(1981), 96–104. http://doi.org/10.1016/j.brainres.2010.03.050

Lin, Y. C., & Shieh, K. K. (2006). Reading a dynamic presentation of Chinese text on a single-line display. *Displays*, *27*(4–5), 145–152. http://doi.org/10.1016/j.displa.2006.04.004

Liu, Z. (2009). The cognitive process of Chinese reflexive processing. *Journal of Chinese Linguistics*, *37*(1), 1–27.

Schmid, M. S., & Dusseldorp, E. (2010). Quantitative analyses in a multivariate study of language attrition: The impact of extralinguistic

factors. *Second Language Research*, *26*(1), 125–160.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, *1*(1), 1–33. http://doi.org/10.1075/lab.1.1.01sor

Sorace, A. (2016). Referring expressions and executive functions in bilingualism. *Linguistic Approaches to Bilingualism*, *6*(5), 669–684. http://doi.org/10.1075/lab.15055.sor

Tsimpli, I., Sorace, a., Heycock, C., & Filiaci, F. (2004). First language attrition and syntactic subjects: A study of Greek and Italian near-native speakers of English. *International Journal of Bilingualism*, *8*(3), 257–277. http://doi.org/10.1177/13670069040080030601

Wang, A. H., & Chen, C. H. (2003). Effects of screen type, Chinese typography, text/background color combination, speed, and jump length for VDT leading display on users' reading performance. *International Journal of Industrial Ergonomics*, *31*(4), 249–261. http://doi.org/10.1016/S0169- 8141(02)00188-9

Building a morphosyntactic lexicon for Serbian using Wiktionary

Aleksandra Miletic
UMR 5263 CLLE, CNRS & University of Toulouse
Maison de la Recherche, 5, allée Antonio Machado
31 000 Toulouse, France
aleksandra.miletic@univ-tlse2.fr

Abstract

Creation of MS lexica often relies on exploiting existing traditional lexical resources or extracting lexical information from raw or annotated corpora. However, this approach is problematic in the case of under-resourced languages like Serbian, for which the starting points for these methods are by definition scarce. An alternative method consists in using new collaborative resources made possible by the evolution of the Internet. The lexicon described in this paper was derived from one such resource: Wiktionary for serbo-croatian. We show that, although the resulting lexicon does not have perfect coverage, this approach allowed us to have a relatively rapid building process resulting in a resource under a non-restrictive license. Some enrichment techniques have also been used in an effort to extend the lexicon coverage.

Keywords: morphosyntactic lexicon, under-resourced languages, Serbian, Wiktionary

1 Introduction

This paper presents the creation of a morphosyntactic lexicon for Serbian derived from several resources: the Wiktionary edition for Serbo-Croatian, a manually POS-tagged corpus, and specialized preposition lists. This work is part of a larger effort to transform ParCoLab (Stosic, 2015), a parallel corpus of Serbian, English, and French, into a syntactic treebank.

English and French already boast a variety of different NLP resources: lexical and morphological resources (e.g. Clément et al., 2004, Romary et al., 2004, Sajous et al., 2013 for French; Fellbaum, 1998, Brierly et al., 2008 for English), POS-tagging methods (e.g. Shen et al., 2007 for English, Denis & Sagot, 2009 for French) and parsers (McDonald et al., 2006, Nivre et al., 2006 for English and French; Urieli 2013 for French). This makes parsing English and French subcorpora of ParCoLab much more straightforward than the annotation of their Serbian counterpart. Serbian is a Slavic language with rich inflectional morphology and flexible word order. This type of languages typically represents a challenge for NLP, and Serbian is not an exception. Despite the recent developments in POS-tagging and lemmatization (Gesmundo & Samardzic, 2012) and first experiments in parsing (Jakovljevic et al., 2014), it can still be considered as an under-resourced language: no training corpora for parsing are available and, to the best of our knowledge, the only freely available morphosyntactically tagged training corpus is still the *cesAna* corpus from the MULTEXT-East project (Krstev et al., 2004a). Although morphological and lexical resources for Serbian are referenced in previous works (Krstev et al., 2004a, 2004b), up to now we have been unable to gain access to them[1]. Consequently, the project of transforming ParCoLab into a syntactically annotated corpus demands an intensive resource-building campaign, in which the creation of a lexicon containing the information relevant to morphosyntactic and syntactic analysis is one of key parts.

The lexicon presented here contains 1 226 638 million wordforms for 117 445 lemmas, corresponding to a total of 3 066 214 unique triples <*wordform, lemma, morphosyntactic description*>. It is thought for NLP applications such as POS-tagging and parsing. It is downloadable under the Creative Commons BY-SA 3.0 license at the following address: http://redac.univ-tlse2.fr/.

2 Related works

Creation of morphosyntactic lexica often relies on exploiting existing lexical resources or extracting lexical information from raw or annotated corpora (cf. Clément et al., 2004, Sagot, 2005). Although previous works reference a morphosyntactic lexicon (Krstev et al., 2004a) and a morphological dictionary in Intex format (Krstev et al., 2004b) for Serbian, we have been unable to gain access to them up to now. Furthermore, they are subject to a no redistribution license, and our goal is to create a freely accessible set of NLP tools for Serbian. As for corpora-based methods, ParCoLab was not large enough for this method to be efficient (1,6M tokens in the Serbian subcorpus) at the beginnig of this work. We therefore looked for alternative methods, which led us to an existing lexical resource for Serbian: Wiktionary.
Wiktionary is a collaborative dictionary launched in 2002. Today it exists in 158 languages, and entries can

[1] It should be noted that a new lexicon for Serbian was published (Ljubesic et al., 2016) after the completion of the work presented here. It will be taken into account in our future work.

contain definitions, as well as information on pronunciation, inflection, semantically related words, translations into other languages, etc. This makes Wiktionary a valuable resource for NLP, but the fact that it is created through crowd-sourcing can put in question the quality of its content and the quality of the resources derived from it. However, several works have shown that resources based on crowd-sourcing can yield results that are competitive or even better than those obtained from resources built by experts (Strube & Ponzetto, 2006, Gabrilovich & Markovitch, 2007, Zesch et al., 2007, Zesch & Gurevych, 2010). Since the first works on Wiktionary in 2008, its suitability for NLP research seems to have become an accepted fact: it has been used to measure semantic relatedness between words (Zesch et al., 2008), create synonymy networks (Navarro et al., 2009), build or enrich ontologies (Meyer & Gurevych, 2012, Pérez et al., 2011), and derive morphosyntactic lexicons (Sajous et al., 2013, Sagot, 2014, Senrich & Kunz, 2014).

As this approach is low-cost compared to a manual creation process, it is especially useful where time and human resources are scarce. It is even more so for low-density languages, for which other starting points for resource derivation can be difficult to find. Both conditions apply to our case. Using Wiktionary also allowed us to have a relatively rapid building process and a resulting resource under a non-restrictive license.

3 Building process

The base for our lexicon was derived from the Wiktionary edition for Serbo-Croatian. Two Wiktionary editions treating Serbian content exist: the Serbo-Croatian one (sh.wiktionary.org) and the Serbian one (sr.wiktionary.org). This seems to be due to extra-linguistic rather than linguistic factors. We chose the Serbo-Croatian edition for its size: 850 000 entries vs. 45 000 in sr.wiktionary.org. Since the lexicon will be used to parse ParCoLab, we focused on getting morphosyntactic information, especially the case, number and gender, as they are essential for syntactic analysis of Serbian.

Wiktionary is made publicly available through periodic XML data dumps. We used the sh.wiktionary.-com dump from October 02 2015. It should be noted that only the macrostructure of the pages is encoded in XML, whereas the page content is rendered in wikicode, a very flexible, under-specified text-based format. Since no systematic description of the wikicode syntax is available, building a parser for wikicode needs to be done through meticulous observation of the pages. As noted in (Navarro et al., 2009, Sajous et al., 2013), the page structure in different language editions varies substantially and a wikicode parser developed for one language cannot be simply transported to a different wiktionary edition. This is why it was necessary to develop a new parser for the Serbo-Croatian Wiktionary.

Another difficulty in parsing Wiktionary stems from the fact that different encoding conventions can coexist within one dump. For example, there are two main page types in the Serbo-Croatian Wiktionary: lemma-based, which gives the complete inflectional paradigm of a lemma in a table (cf. Figure 1), and wordform-based, in which the entry is an inflected wordform, for which all the possible morphosyntactic interpretations are given (cf. Figure 2).

In the first format, the morphosyntactic properties of each form are either given through codes or need to be deduced from the position of the form in the table (typically the case of nouns, cf. Figure 1). This is possible because the tables follow the generally accepted case ordering for Serbian. However, some articles were found where the instrumental and the locative forms switched places. In order to ensure the correct case information, our parser performs a rule-based check to verify that the supposed case corresponds to the wordform ending.

```
==== Deklinacija ====
{{sh-imenica-deklinacija2
|jezik|jezici
|jezika|jezika
|jeziku|jezicima
|jezik|jezike
|jeziče|jezici
|jeziku|jezicima
|jezikom|jezicima
}}
```

Figure 1: Lemma-based article

In the wordform-based format, the information is given in the form of textual descriptions (cf. Figure 2). The order of the elements is not fixed, and some pieces of information can be missing. The parser needed to be flexible enough to manage this diversity in order to extract as much data as possible.

```
=== Flektirani oblici ===
'''guvernerskim'''

# instrumental množine ženskog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# lokativ množine ženskog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# dativ množine muškog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# instrumental množine muškog roda pozitiva
određenog vida pridjeva
```

Figure 2: Form-based articles

Other resources at our disposition were used to improve the coverage. 107 prepositions were imported from manually created lists resulting from previous theoretic work on spatial relations in Serbian (Stosic, 2001). 76 additional prepositions, 43 conjunctions, 33 interjections and 868 adverbs were extracted from the manually POS-tagged part of the ParCoLab corpus (Miletic, 2013) and integrated in the lexicon.

4 Quality of the lexicon

The lexicon presented here contains 1 226 638 million wordforms for 117 445 lemmas, corresponding to a total of 3 066 214 unique triples <*wordform, lemma, morphosyntactic description*>. For comparison, GLAFF, a lexicon for French derived from the wiktionary, contains 1 425 848 wordforms corresponding to 186 082 lemmas. Our lexicon is in plain text format as illustrated in Figure 3. The first column contains the inflected wordform followed by one or more morphosyntactic descriptions (MSDs). The structure of the MSDs is POS-specific, but in each case the first slot indicates the POS, and the last one the lemma, while the intervening slots encode values of different morphosyntactic properties.

```
trag    N_m_nom_sg_trag N_m_acc_sg_trag
traga    V_Present_3_sg_0_tragati
N_m_gen_sg_trag
tragah  V_Imparfait_1_sg_0_tragati
tragahu V_Imparfait_3_pl_0_tragati
tragaj  V_Imperatif_2_sg_0_tragati
```

Figure 3: Lexicon format

The values of the MS properties are given in a relatively explicit format in order to facilitate the manual verification. Another version of the lexicon with MSDs in the more standard MULTEXT-East format (Krstev et al., 2004a) will also be made available.
The MS properties for inflected categories are given in Table 1.

POS	MS properties encoded in lexicon
Verb	verb form, person, number, gender
Noun	gender, case, number
Pronoun	case, number, gender
Adjective	case, number, gender, degree of comparison
Adverb	degree of comparison (if applicable)

Table 1: Morphosyntactic properties for inflected classes

In order to evaluate the lexicon, we calculated its coverage over a portion of ParCoLab. The texts used in the test come from 3 contemporary novels, containing 150 000 tokens equivalent to 28 980 unique wordforms. The coverage was calculated for all wordforms, and then for those appearing at least 2, 5 and 10 times in the subcorpus (cf. Table 2). Eliminating wordforms that occur only once improves the coverage for 4.7%, but these wordforms constitute more than 50% of the identified unique wordforms (cf. number of unique wordforms for thresholds 1 and 2). This is probably due to the relatively small size of the subcorpus used for coverage calculation. In order to have more reliable results, the test will be repeated with a larger portion of ParCoLab.

frequency treshold	# of unique wordforms	Found in lexicon	Coverage
1	28 980	20 808	71.80%
2	10 630	8 136	76.53%
5	2 946	2 328	79.02%
10	1 241	990	79.77%

Table 2: Lexicon coverage

These results also show that although the lexicon is a solid starting point, the resource needs to be developed further. One of the possibilities is to develop a parser for sr.wiktionary.com, which could contain valuable additions to the exissting ressource. We are also considering the possibility of ranking the wordforms found in ParCoLab but not in the lexicon by frequency and adding the most frequent ones manually.

We also performed a quantitative analysis of the lexicon, which gave us insight into ambiguity of Serbian. For 1.2 million wordforms in the lexicon, there are more than 2.5 million MSDs (2.1 MSD per wordform). 727 000 wordforms (60%) are ambiguous. Furthermore, the number of MSDs per wordform can be very high: more than 37 000 wordforms have 10 or more associated MSDs, with 5 wordforms reaching a maximum of 43 MSDs. Although wordform ambiguity in Serbian is intuitively high, the existence of wordforms with 15 or more MSDs seems noteworthy. A manual evaluation of these highly ambiguous wordforms will be performed to exclude errors due to the extraction method or to the quality of the source articles.

These results incited us to try to identify different types of ambiguity in the lexicon. We distinguished 4 categories: i) wordforms corresponding to different lemmas belonging to different POS categories (cf. *krilo*, which can be a nominative/accusative singular of the noun *krilo* ‘lap’, or neuter singular of the past participle of the verb *kriti* ‘to hide’), ii) those corresponding to lemmas having the same form but belonging to different POS categories (cf. *blizu*, which can be a preposition ‘near’ or an adverb ‘nearby’), iii) those corresponding to different lemmas, but having the same POS category (cf. *vrata*, genitive singular of the noun *vrat* ‘neck’, or nominative/accusative plural of the noun *vrata* ‘door’), and iv) ambiguous forms belonging to the paradigm of the same lemma (cf. *jastucima*, which can be dative, instrumental or locative plural of the noun *jastuk* ‘pillow’). The results of this

analysis show that 95% of the ambiguous wordforms belong to the last category (cf. Table 2). This indicates that a large part of ambiguity in Serbian is due to the syncretism in inflectional paradigms.

	# of word-forms	% of all ambiguous wordforms
Ambiguous POS and lemma	15 496	2.13%
Ambiguous POS, unambiguous lemma	303	0.04%
Unambiguous POS, ambiguous lemma	19 822	2.72%
Unambiguous POS and lemma, ambiguous MS properties	691 814	95.10%

Table 3: Ambiguity anlaysis

5 Conclusions and future work

This work presents a new lexicon for Serbian containing 1.2 million wordforms corresponding to over 126 000 lemmas. The ressource was created through the use of complementary resources: the greater part of the content was derived from Wiktionary for Serbo-Croatian and subsequently completed with closed-class words from manually created lists and a manually POS-tagged subcorpus. This approach allowed us to reach a solid coverage on a contemporary literary corpus, but the results also showed that there is still room for improvement. For this, we are considering two main approaches: we will explore the possibility to exploit the Serbian edition of Wiktionary (sr.wiktionary.com), which could contain valuable additions to the existing resource. We will also test a semi-manual enrichment process based on frequency lists of wordforms found in ParCoLab, but missing from the lexicon.

References

Clément, L., Lang, B., and Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC2004), p. 1841–1844, Lisbon, Portugal.

Brierley, C. and E. Atwell. (2008). ProPOSEL: a Prosody and POS English Lexicon for Language Engineering. In *Proceedings of LREC'08 Language Resources and Evaluation Conference*, Marrakech, Morocco. May 2008.

Denis, P., & Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation.* Hong Kong.

Fellbaum, C., editor. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Gabrilovich, E., and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1606–11, Hyderabad, India.

Gesmundo, A., & Samardžić, T. (2012). Lemmatising Serbian as a category tagging task with bidirectional sequence classification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation.* Istanbul.

Jakovljević, B., Kovačević, A., Sečujski, M., & Marković, M. (2014). A Dependency Treebank for Serbian: Initial Experiments. *Speech and Computer Lecture Notes in Computer Science, 8773*, pp. 42-49.

Krstev, C., Vitas, D., & Erjavec, T. (2004a). MULTEXT-East resources for Serbian. In *Proceedings of 7th Informational Society - Language Technologies Conference*, pp. 108-114. Ljubljana.

Krstev, C., Vitas, D., Stanković, R., Obradović, I., & Pavlović-Lažetić, G. (2004b). Combining heterogeneous lexical resources. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, pp. 1103-1106.

Ljubešić, N., Klubička, F., Agić, Ž., Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, p. 4264-4270.

McDonald, R., Lemran, K., & Pereira, F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X).*

Meyer, C. M. and Gurevych, I. (2012). OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In Pazienza, M. T. and Stellato, A., editors, *Semi-Automatic Ontology development: Processes and Resources*, chapter 6, pages 131–161. IGI Global, Hershey, PA, USA.

Miletic, A. (2013). Annotation semi-automatique en parties du discours d'un corpus littéraire

serbe. *Mémoire de Master*, Université Charles de Gaulle Lille 3.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., and Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Singapore.

Nivre, J., Hall, J., & Nilsson, J. (2006). MaltParser A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*.

Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pp. 703–717, Lisbon, Portugal.

Romary, L., Salmon-Alt, S., and Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. Zock, M. and Saint-Dizier, P., editors, *COLING 2004 Enhancing and using electronic dictionaries*, pp. 22–28, Geneva, Switzerland.

Sagot, B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Text, Speech and Dialogue*, pp. 156-163, Springer, Berlin Heidelberg.

Sagot, B. (2014). DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland.

Sajous, F., Hathout, N., and Calderone, B. (2013a). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pp. 285–298, Les Sables d'Olonne, France.

Sennrich, R., & Kunz, B. (2014, May). Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Shen, L., Satta, G., & al. (2007). Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 760-767. Prague.

Stosic, D. (2001). Le rôle des préfixes dans l'expression des relations spatiales. Eléments d'analyse à partir des données du serbo-croate et du français. *Cahiers de Grammaire 26*, p. 207-228.

Stosic, D. (2015). ParCoLab (beta), A Parallel Corpus of French, Serbian and English. Toulouse, France: CLLE-ERSS, CNRS & Université de Toulouse 2. (http://parcolab.univ-tlse2.fr)

Strube, M., and Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on ArtificialIntelligence (AAAI-06)*, pp. 1419–24, Boston, MA.

Urieli, A. (2013). *Analyse syntaxique robuste du français : concilier méthodes statistiques et connaissances linguistiques dans l'outil Talismane.* PhD thesis. Université Toulouse II le Mirail.

Zesch, T., and Gurevych, I. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pp. 1–8, Rochester, NY. Association for Computational Linguistics.

Zesch, T. et Gurevych, I. (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering*, 16(01):25–59.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008)*, Marrakech, Morocco

Compass: a parallel French-Russian corpus enriched with morpho-syntactic annotation

Olga Kataeva
L'Institut Catholique de Toulouse
31 Rue de la Fonderie
31000 Toulouse
olga_kataeva@yahoo.com

Elena Manishina
IRIT (UT3)
18 Route de Narbonne
F-31062 Toulouse
elena.manishina@irit.fr

Abstract

Despite the existence of multiple bilingual resources nowadays, parallel corpora for rare language couples, like Russian-French remain scarce. The existing corpora represent in their majority parallel texts, aligned at the sentence level without any form of parallel annotation (morpho-syntactic, semantic, pragmatic, etc.). Automatic annotation tools traditionally used to obtain morpho-syntactic information are error-prone and often require manual correction/validation.

In this paper we present Compass : a new bilingual French-Russian resource annotated with morpho-syntactic information on both sides. It represents a comprehensive resource that can be used to perform comparative linguistic analysis and to build statistical machine translation models. Furthermore each side of the corpus can be used separately as a monolingual resource to train statistical syntactic parsers and morphological analyzers.

Keywords : parallel corpus, morpho-syntactic analysis, corpus annotation

1 Introduction

Despite the existence of multiple bilingual resources nowadays, parallel corpora for rare language couples, like Russian-French remain scarce. The existing corpora represent in their majority parallel texts, aligned at the sentence level ; to our knowledge none of the existing bilingual resources contains any form of annotation (morpho-syntactic, semantic, pragmatic, etc.) which makes it hard to build linguistically enriched translation models (factored models, syntactic models, etc.) using these corpora. Automatic annotation tools traditionally used to obtain morpho-syntactic information are error-prone and often require manual correction/validation ; in many cases they also require manually built monolingual training resources.

In this paper we present Compass : a new bilingual French-Russian resource annotated with morpho-syntactic information on both sides. It represents a collection of sentence-aligned bi-texts derived from press releases of the Council of Europe [1] with alignment validation and morpho-syntactic annotation performed by language experts.

Both sides of the corpus are manually annotated with morpho-syntactic categories (see section 4) using the same tagset. The common set of morpho-syntactic categories is the result of elaboration and analysis of the existing monolingual resources and annotation guidelines ; the objective was to facilitate the comparative analysis and parallel processing of any sort.

Compass is a comprehensive resource that can be used for different purposes : to teach French or Russian to advanced language students or future translators, to perform comparative linguistic analysis, to build statistical machine translation (SMT) models, etc. Each side of the corpus can be used separately as a monolingual resource to train statistical syntactic parsers and morphological analyzers.

The paper is structured as follows : in section 2 we give an overview of the existing bi-lingual (2.1) and monolingual (2.2) resources for French and Russian ; section 3 discusses the initial corpus collection : data selection and normalization (3.1) and alignment (3.2) ; in section 4 we present our annotation scheme, specifically lexical categories (4.1) and syntactic annotation (4.2) ; finally, we conclude the paper with a brief discussion in 5.

1. https ://wcd.coe.int/

2 Background

With the advent of statistical methods in machine translation and morpho-syntactic analysis grew the interest in building parallel and monolingual resources with various kinds of morpho-syntactic annotations. Today there exists a significant number of monolingual and bilingual French-Russian resources. In this section we will present the most widely used ones as well as the closest to our corpus in nature and objective.

One of the major limitations of many of the existing annotated resources (RUSCORPORA, Frantext) is the 'on-line' consultation of the corpus : neither raw (unannotated) nor annotated data is available for download. The obligatory access to complete raw texts is one of the principles for corpus creation proposed by Sinclair (see section 3.1).

2.1 Bilingual corpora

MultiUN is a collection of translated documents from the United Nations ([Eisele and Chen, 2010]) proceedings ; it contains 79K documents with 13M sentences and currently represents one of the largest bilingual French-Russian corpus.

Open Subtitles is a collection of translated movie subtitles [2] [Lison and Tiedemann, 2016]. It contains 13.7M sentences. GNOME is a parallel corpus of GNOME localization files [Tiedemann, 2012] with 0.8M sentences.

A parallel corpus of News Commentaries is provided by WMT for training SMT models [3]. The size of the latest edition (11th) is 0.2M sentences.

Russian National Corpus (RUSCORPORA) contains a subcorpus of parallel Russian-French fiction texts (100K) [4]. The specificity of this corpus is that it contains translation variants for each given sentence on both sides. Thus it may be considered a multi-variant parallel corpus.

EMOBASE is a multilingual database from EMOLEX project [5] which contains comparable corpora (news and fiction) in French, English, German, Spanish and Russian. The French-Russian part contains 17 texts with 1,3M words in total.

A parallel corpus of XIX century has 13,7K and 15K lines on the French and Russian sides respectively [6].

These corpora represent a great source for training SMT models, but none of them contains any kind of annotation ; they all represent a plain text on both sides (languages) aligned at the sentence level.

2.2 Monolingual corpora

Among the monolingual Russian language resources, the biggest and the most widely used is The Russian National Corpus (100M words) [7] ; the corpus is annotated with morphological (word) categories ; it also contains a subcorpus (30K) which has morpho-syntactic annotations (dependency trees). Three other types of annotation are metatextual, word stress and semantic ones. Other significant monolingual corpora include Russian Internet Corpus (90M words), a corpus of Russian newspapers (78M words) and the Russian Standard - a corpus of modern Russian fiction with manual disambiguation of morphological categories (1.6M words).

As for French, the biggest and the most well-known annotated resource today is the French Treebank [Abeillé et al., 2003]. Another major corpus is Frantext which is the collection of texts spanning from X to XXI century and having around 300 million words. Other resources include the bilingual annotated English-French International Telecommunications Union corpus hosted by The Corpus Resources And Terminology Extraction project (2M tokens with human-edited morpho-syntactic annotations), Sequoia [Candito and Seddah, 2012] - a 3,1K-sentence corpus annotated with constituency trees and later also with deep syntactic dependency trees, MULTEXT JOC Corpus (appr. 200K words grammatically tagged and manually checked)[Véronis and Khouri, 1995] and PAROLE (48,4K words annotated with morphological and syntactic information.)

3 The corpus

In this section we describe the data collection procedure. It includes document selection, text normalization and sentence alignment. We constructed our corpus in accordance with the theoretical foundations for corpus building laid out by the French (Condamines A., Habert B.), Russian (Dobrovolski D., Ploungyan V.) and British

2. http ://www.opensubtitles.org/
3. The source is taken from CASMACAT : http ://www.casmacat.eu/corpus/news-commentary.html
4. http ://ruscorpora.ru/search-para-fr.html
5. www.emolex.eu
6. http ://nevmenandr.net/fr/index.php ?go=head
7. http ://ruscorpora.ru

[s16] Хальвдан Скард призывает международные организации воспользоваться потенциалом примирения , имеющимся у местных властей на Ближнем Востоке

[s17] Halvdan Skard encourage les organisations internationales à utiliser le potentiel de réconciliation des pouvoirs territoriaux au Moyen Orient

FIGURE 1: The output of Alinéa : sentence-level alignment

corpus linguists (Halliday M.A.K., Sinclair J.).

3.1 Data collection and normalization

To constitute the corpus we resorted to the website of the Council of Europe (EC), which contains documents drafted in the languages of the countries members.

The multilingual content available on the EC website (press releases, thematic files, official documents) allows for building extensive bilingual and monolingual corpora. Specifically as the covered topics/areas include various spheres : social, political, economic, etc. The site is constantly updated with new material.

To constitute the corpus we follow the guidelines proposed by John Sinclair in his work "EAGLES. Preliminary recommendations on Corpus Typology" [Eag-Tcwg-Ctyp, 1996], specifically the following criteria :

1. Using complete documents without cutting and/or reshuffling
2. The corpus must contain parallel texts (not comparable or other)
3. The translation is performed from French to Russian
4. The time frame is well defined : the corpus contains texts representing the language between 1950 till 2014 (1st release of the corpus).
5. The corpus must be aligned at the sentence level
6. 'Representativeness' of the corpus is highly desirable The notion of '*representativeness*' for a bilingual corpus is defined here as containing equal (or close) proportions of texts from different genres and covering different topics.

Not all the language versions are present for each specific document. So the first step in retrieving a given press release/document is to determine the presence of a French and Russian translations ; if both versions are available, the document is downloaded. The next step is to perform automatic sentence-level alignment, which is manually verified and corrected in case of mismatch.

3.2 Alignment

To perform sentence alignment we used Alinéa tool [8] developed by Olivier Kraif. This software uses statistical and linguistic features to find an optimal sentence segmentation and alignment. We performed a number of tests with other alignment tools, including UNITEX [9] but Alinéa turned out to have the highest precision.

It is important to have texts on both sides with similar sentence segmentation before feeding it to the alignment software. This is generally the case in press releases and official documents which have similar document structures in both Russian and French. The situation is different with technical documentation which is distributed in PDF format (compared to standard HTML in case of official documents) : the alignment is preceded by extraction of pure text from PDF, which does not result in similar document structures for the two languages, specifically in case of complex textual entities like tables. Treating such documents require an extensive manual alignment preprocessing which is the reason why the technical documentation subcorpus is considerably smaller than the other two parts in our corpus.

The automatic alignment is performed in three steps : extraction of anchor points, phrasal alignment (calculation of the best alignment path) and extraction of lexical correspondences. The automatic alignment is then manually verified and corrected.

3.3 Corpus statistics

The corpus statistics is presented in Table 1. As of today, the corpus has 523701 words on the French side and 414146 words on the Russian side ; 56% from the official documents of the European council (conventions, additional protocols, agreements, recommendations, resolutions, decla-

8. http ://olivier.kraif.u-grenoble3.fr
9. www-igm.univ-mlv.fr/ unitex/

Corpus	Words fr	Words ru
Legal documents	293047	235772
Press releases	211395	161677
Technical documentation	19259	16697
Total	**523701**	**414146**

TABLE 1: Compass corpus statistics

rations, statutes, charts) ; 40% are press releases (2006-2007) and thematic files of the EC website and 4% are technical documentation.

4 Annotation

For part-of-speech tagging we use the annotation specifications and tagsets common for most morphological analyzers (RUSCORPORA, simplified TreeTagger tagset [Schmid, 1995], etc.) To annotate the French side of the corpus we followed the guidelines for morpho-syntactic annotation of the French Treebank [10]. For the Russian side we resort to the annotation description provided on the website of RUSCORPORA (Russian National Corpus). [11].

4.1 Word categories (tagset)

For the Russian side of the corpus we use the tagset defined for the RUSCORPORA as a base. We modify the basic tagset by splitting a generic N (nouns) tag into NC (common noun) and NP (proper noun) and including ET (foreign word) tag from the Treebank tagset. For the french side we opt for the tagset elaborated for the French Treebank. Here again we slightly modify the tagset to include the following tags from RUSCORPORA : NUM (numeral), A-NUM (numeral adjective) and PART (particle). Our final tagset is presented in Table 2 : here we first outline the tags common for both languages in a joint common tagset ; then the language-specific tags are listed for both sides of the corpus.

There is a number of other differences between our tagset and the Treebank tagsets (apart additional tags). In the Treebank most typographical signs (including %, numbers and abbreviations) are assigned an N tag (common noun). We use a specific marker SIGN for mathematical symbols, currencies, etc and an ABBR tag for abbreviations.

10. http ://www.llf.cnrs.fr/sites/sandbox.linguist.univ-parisdiderot.fr/files/statiques/french_treebank/guide-annot.pdf

11. http ://ruscorpora.ru/en/corpora-morph.html

Tag	Category
ABBR	abbreviation
A	adjective
Adv	adverb
Conj	conjunction
NUM	numeral
A-NUM	numeral adjective
V	verb
CS	conjunction
ET	foreign word
I	interjection
NC	common noun
NP	proper noun
P	preposition
PRO	strong pronoun
PART — particle PUNCT	punctuation mark
SIGN	symbol
Additional tags (French) :	
Cl	weak clitic pronoun
D	determiner
PREF	prefix
Additional tags (Russian) :	
A-PRO	adjectival pronoun
ADV-PRO	adverbial pronoun
PRAEDIC	predicative
PARENTH	parenthesis
PRAEDIC-PRO	predicative pronoun

TABLE 2: Lexical tagset

Also we do not distinguish between strong and weak punctuation markers like it is the case in the Treebank - all punctuation marks are tagged with PUNKT. But we do keep the granularity in pronouns borrowed from the RUSCORPORA (and absent from the Treebank) since we think it reflects well the realities of the Russian language.

4.2 Syntactic annotation scheme

To define the protocol for the syntactic annotation we resort to the guidelines for annotation of the French Treebank. Here again we modify the initial phrasal tagset in order to make it suitable for both the Russian and the French sides of the corpus (Table 3).

We remove COORD tag (coordinated phrases)

```
<SENT>
  <NP>Elle/PRO</NP>
  <VP>vise/VN
    <PP>à/PP
      <VPinf>intensifier/VN
        <NP>les/D efforts/N
          <PP>de/P
            <NP>ses/PRO Etats/N membres/N</NP>
          </PP>
          <PP>dans/P
            <NP>la/D prévention/N <PP>du/P terrorisme/N</PP><NP>
          </PP>
        </NP>
      </VPinf>
    </PP>
  </VP>
./PUNKT
</SENT>
```

(a) French

```
<SENT>
  <NP>Она/PRO</NP>
  <VP>направлена/PPT
    <PP>на/PP
      <NP>активизацию/N
        <AP>антитеррористических/A усилий/N
          <NP>государств/N -/PUNKT
            <NP>членов/N
              <NP>Совета/NP Европы/NP</NP>
            </NP>
          </NP>
        </AP>
      </NP>
    </PP>
  </VP>
./PUNKT
</SENT>
```

(b) Russian

FIGURE 2: Parse trees for a sentence : "Elle vise à intensifier les efforts de ses Etats membres dans la prévention du terrorisme."

Tag	Category
AP	adjectival phrases
AdP	adverbial phrases
NP	noun phrases
PP	prepositional phrases
VN	verbal nucleus
VPinf	infinitive clauses
VPpart	nonfinite clauses
SENT	sentences

TABLE 3: Phrasal tagset

and the tags reflecting the distinction between different types of finite clauses (as they are defined in the Treebank). We only (implicitly) keep the distinction between finite and non-finite clauses by preserving the non-finite clause tag (VPpart). Figure 2 depicts the parse trees for a sentence "Elle vise à intensifier les efforts de ses Etats membres dans la prévention du terrorisme." and it's equivalent in Russian.

5 Conclusion

In this paper we presented a new parallel corpus for French-Russian language couple enriched with manual morpho-syntactic annotation on both sides. There are many possible applications for the corpus : building grammatically enriched statistical machine translation models, train statistical syntactic parsers and morphological analyzers, performing different kinds of morphological and/or syntactic analysis, etc.

The corpus is constantly growing. The new version Compass-v2.0 is scheduled for 2017, with additional 2K parallel sentences annotated with morpho-syntactic information. A part of the version 1.0 of Compass is freely available on the corpus website [12]. We will continue growing our corpus and improve its representativeness. Our goal is an equal distribution with roughly 25% of each of the following categories : press releases, official documents, technical documentation and thematic files. We also plan to extend the corpus with texts translated from Russian to French (as opposed to French-Russian translations which currently represent the majority of the texts in the corpus).

References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In *Treebanks*, pages 165–187. Springer.

Candito, M. and Seddah, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN'2012*.

Eag-Tcwg-Ctyp, E. D. (1996). Eagles preliminary recommendations on corpus typology.

Eisele, A. and Chen, Y. (2010). Multiun : A multilingual corpus from united nation documents. In Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, K., and Chair), N. C. C., editors, *Proceedings*

12. http ://www.tageater.com/Compass

of the Seventh conference on International Language Resources and Evaluation, pages 2868–2872. European Language Resources Association (ELRA).

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016 : Extracting large parallel corpora from movie and tv subtitles.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA).

Véronis, J. and Khouri, L. (1995). Étiquetage grammatical multilingue : le projet multext. *Traitement Automatique des Langues*, 36(1/2) :233–248.

« Cuisinez chic » : les emplois adverbiaux de l'adjectif en français

COIFFET Benoit
Université Toulouse Jean Jaurès
Laboratoire CLLE-ERSS
Benoit.coeffet@icloud.com

Résumé

Le phénomène des emplois adverbiaux de l'adjectif est connu dans la littérature linguistique, mais les études approfondies à son sujet restent marginales. Après avoir montré que ces emplois sont atypiques, mais en pleine expansion, nous approfondissons les pistes explorées précédemment dans Grundt (1972), puis dans Noailly (1994). Ces deux auteurs, travaillant sur des corpus relativement restreints et littéraires, limitent la portée sémantique de l'adjectif à droite d'un verbe à deux cas principaux : soit l'adjectif ne porte que sur un objet non exprimé à droite d'un verbe transitif, soit il caractérise la manière dont le procès a été réalisé. En adoptant une démarche se situant à l'interface entre la syntaxe et la sémantique, nous montrons qu'il existe des cas intermédiaires dans lesquels un adjectif invarié à droite d'un verbe transitif peut à la fois caractériser un objet non exprimé et spécifier la manière dont le procès a été réalisé.

Mots-clés : adjectif – emploi adverbial – manière.

1 Introduction

On observe de nos jours une profusion d'emplois où un adjectif invarié accompagne le verbe :

(1) Ce soir, j'ai cuisiné *chinois*. (Internet)
(2) L'Île-de-France : 1ère collectivité à emprunter *responsable*. (Internet)
(3) Mâchez *danois* ! (Publicité stimorol)
(4) Manger *bio*, c'est sûrement meilleur, mais nettement plus cher ! (Entendu, conversation)

Ces emplois se distinguent nettement d'énoncés comme *Il est parti déçu*, où l'adjectif *déçu* est en emploi attributif, ce qui se marque par son accord avec le sujet dont il prédique une propriété concomitante au moment de l'action exprimée par le verbe. Les emplois (1) - (4) sont à considérer comme des emplois atypiques de l'adjectif, emplois dans lesquels on a l'impression qu'il occupe une position normalement réservée à l'adverbe.

Bien qu'identifiés comme « emplois adverbiaux » de l'adjectif dans la littérature, les études sur la question restent rares. En effet, à la suite de Grundt (1972) qui, dans son étude fondatrice en la matière, propose une approche systématique des emplois adverbiaux de l'adjectif en français, de nombreux linguistes se sont intéressés à la construction [Verbe + Adjectif Invarié] (désormais [V+Adj.Inv.]).

D'aucuns comme Noailly (1994 : 105) ont pu voir dans cette construction un reliquat de l'ancien français dans des tours comme *il a payé cher sa voiture, elle hache menu ses carottes* ; ces exemples sont caractérisés par leur possibilité de voir apparaître à la fois l'objet syntaxique du verbe (*sa voiture, ses carottes*) et un adjectif invarié intercalé entre le verbe et le complément d'objet.

À côté de ces emplois, on trouve des constructions dans lesquelles l'adjectif invarié apparaît, sans objet réalisé lexicalement à droite de V dans le cas de verbes transitifs comme *écrire, cuisiner* (on trouve par exemple dans l'article de Noailly *Écrire gros et lisiblement ;*

Cuisinez transparent), et des cas où [Adj.Inv.] est à droite d'un verbe intransitif (*rouler utile*).

Grundt (1972), et les auteurs qui se sont penchés à sa suite sur la question de l'emploi adverbial de l'adjectif[1], étudient les effets de sens subtils mais bien réels qui sont en jeu dans la construction [V+Adj.Inv.], sur la base de la distinction guillaumienne de l'incidence et de la portée, ainsi reprises par Guimier (1996) : l'incidence est le support syntaxique de l'adverbe, c'est-à-dire l'unité linguistique à laquelle il est rattaché ; la portée constitue pour sa part la *référence sémantique* liée à l'utilisation de l'adverbe, c'est-à-dire l'élément duquel on prédique une information à travers son utilisation. Par conséquent, un adverbe peut être incident à un verbe et porter sémantiquement sur d'autres éléments de la structure argumentale de ce dernier, tels que l'objet, le résultat, ou encore le sujet.

L'analyse de Grundt ouvre ainsi la voie à des observations très fines des effets de sens qui sont en jeu dans la construction [V+Adj.Inv.], avec en toile de fond l'idée qu'il existe une concurrence entre l'adjectif dans ce type d'emploi et l'adverbe en *-ment*, lui-même dérivé d'un adjectif. Dans ce duo [Adj.Inv]/[Adv.-*ment*], chaque forme semble ainsi se réserver une part d'effets de sens qui lui est propre à droite de V, et qu'il faudra mettre en évidence notamment à l'aide de tests syntaxiques.

Il faudrait toutefois noter que Grundt travaille sur un corpus d'exemples essentiellement littéraires ou présentant un certain nombre de traits de figement, ce qui a des répercussions essentielles sur l'analyse et le classement qu'il propose, éléments d'analyse qu'on retrouve dans la plupart des études qui lui font suite. Dans les grandes lignes, les critiques s'accordent à répartir les cas en deux tendances exclusives l'une de l'autre[2] : si [Adj.Inv.] à droite de V est à chaque fois *incident* au verbe, *soit* il porte sur la manière d'effectuer le procès exprimé par le verbe (interprétation 'manière'), *soit* il porte sur un objet syntaxique de Vtr qui n'est pas lexicalisé dans l'énoncé (interprétation 'objet').

Bien que Grundt apporte certaines nuances, par exemple en distinguant comme le fait Moignet l'incidence « précoce » (sur le sujet et sur le verbe) de l'incidence « tardive » (sur le verbe uniquement), les cas qu'il étudie montrent une véritable étanchéité d'interprétation : [Adj.Inv.] ne peut entrer que dans une seule interprétation à la fois. On retrouve cette dualité dans la très grande majorité des études de l'emploi adverbial de l'adjectif.

À l'instar de l'étude de Noailly, qui observe dans le courant des années 90 (1994) qu'il s'agit là « d'un modèle syntaxique dont la productivité *n'est pas énorme*[3], mais progresse toutefois, en partie par le jeu des slogans publicitaires et autres formules destinées à attirer l'attention de tout lecteur », nous avons pu constituer, une vingtaine d'années plus tard, un corpus nettement plus étendu que le sien, puisque nous recensons quelque 506 énoncés basés sur une cinquantaine de verbes et une centaine d'adjectifs, contre seulement une vingtaine d'exemples collectés au cours de nos lectures critiques, le tout collecté sur la base de trois sources : la base de données FRANTEXT, le moteur de recherche GOOGLE, et enfin un ensemble d'exemples personnels obtenus au fil de conversations, ou entendus (métro, radio…)[4].

Mais le corollaire d'une telle extension des données est une nécessaire évolution du fil interprétatif de la construction : où Noailly et les autres linguistes ne fondaient leur analyse que sur un nombre assez restreint d'exemples, nous disposons de notre côté d'une base de données conséquente, grâce à laquelle nous pouvons proposer une analyse plus nuancée de l'alternance entre les interprétations 'manière' et 'objet' en jeu dans la construction [V+Adj.Inv.].

L'objectif de ce travail est donc de montrer que la frontière entre ces deux types d'emplois n'est pas nécessairement aussi nette qu'il n'y paraissait au départ. Nous concentrerons notre analyse sur des verbes transitifs et montrerons

[1] Nous pensons par exemple à l'étude des degrés de figement dans la construction [V+Adj.Inv.] de Guimier & Oueslati (2006) ou encore à la réflexion stimulante dans le cadre de la grammaire HPSG d'Abeillé & Godard (2004), qui essaient d'attribuer un poids selon l'opposition léger/non léger, dans cette même construction.

[2] Certaines nuances existent chez ces auteurs, que nous n'avons pas la place de développer dans le cadre de ce travail, mais la répartition, elle, semble partagée unanimement.

[3] Nous soulignons.

[4] Notre corpus a été constitué entre 2010 et 2014. Frantext a été exploité sans limitation dans le temps. Le moteur de recherche GOOGLE et les données personnelles ont été utilisés entre 2010 et 2014. Il est à noter que l'essentiel des énoncés intégrant une séquence [V+Adj.Inv.] obtenus via Frantext est concentré sur le XX°S.

que l'adjectif invarié à droite de Vtr peut caractériser *à la fois* la manière *et* un objet non exprimé ; c'est dire ainsi que notre travail se situe au cœur de l'interface entre syntaxe et sémantique puisque notre objet d'étude consiste en une seule et unique construction syntaxique ([V+Adj.Inv.]), à l'intérieur de laquelle se nouent des relations sémantiques variées qui dépendent de l'interaction directe entre un verbe et un adjectif.

Nous reprenons dans un premier temps les prémisses d'analyse proposées dans Noailly (1994) pour, dans un deuxième temps, exposer un certain nombre de limites ; enfin dans un dernier temps, nous appuyant sur une base définitionnelle de la manière empruntée à Moline & Stosic (2016) et sur les outils d'analyse de l'interaction entre noyau verbal et circonstants exposés par Melis (1983), nous essaierons de montrer qu'une portée de [Adj.Inv.] simultanée sur le verbe et sur l'objet est possible dans le cadre de cette construction.

2 Études fondatrices : Grundt (1972), Noailly (1994) *et sqq.*

Comme nous l'avons dit, Grundt et, dans sa lignée, la plupart des auteurs qui se sont intéressés à la construction [V+Adj.Inv.] sont d'accord pour opposer les fonctionnements 'manière' (rénover *écologique*, rouler *électrique*) et 'objet' (tricoter *chaud*, acheter *français*) de l'adjectif à droite de V.

Le premier mode de fonctionnement est dit être très proche de celui des adverbes en *–ment*, « au point d'incidence près » pour reprendre la terminologie guillaumienne : selon Moignet (1963), on peut effectivement opposer les adverbes en *–ment* selon qu'ils ont une incidence sujet-verbe (*Pierre écoute attentivement*) ou une incidence strictement verbale (*Pierre attend vainement*). À sa suite, Noailly considère que les adjectifs en emploi adverbial sont à ranger dans la deuxième catégorie. Dans les exemples suivants que nous reprenons à cet auteur, [Adj.Inv.] est donc analysé comme un complément de manière[5] :

(5) Vous toussez *gras* ? (Entendu en pharmacie, 1984)

(6) Je voulais savoir s'il était facile ou difficile de danser *contemporain* sur Mozart. (FI, 13/4/94 à 13H55)

Toujours dans les emplois 'manière', Noailly propose d'intégrer les cas où l'Adj.Inv. permet de caractériser les « modalités circonstancielles de la réalisation (du procès)[6] » (p.107) :

(7) Vous vous rasez *électrique* ? (Corpus Moignet)
(8) Cuisinez *transparent*. (Maison de Marie-Claire, n°166)
(9) Dormez *ferme*. (publicité des matelas Lattoflex)

Nous pouvons d'ores-et-déjà constater que, hormis le cas de 'cuisiner', les exemples retenus par l'auteur pour l'interprétation 'manière' sont construits sur la base de verbes intransitifs et pronominaux, ce qui n'est pas sans conséquence dans la répartition des occurrences comme nous l'avons déjà observé.

À l'opposé de ces emplois 'manière' se trouvent les emplois 'objet'. Noailly fait appel à l'analyse proposée chez Riegel, Pellat & Rioul (1994) qui voient dans cette construction la combinaison de deux autres modèles syntaxiques, l'objet interne et la construction à attribut de l'objet. Elle cite : « L'adjectif, dans ce type de construction, caractérise le verbe (mais indirectement, par l'intermédiaire d'un objet générique non exprimé) et il demeure invariable faute d'un objet lexical réalisé avec lequel s'accorder ». Elle propose d'illustrer ces emplois par les exemples suivants qui, selon elle, permettent de définir contrastivement les deux interprétations de [Adj.Inv.] :
(10) (a) Achetez *beau*.
(b) Achetez *réfléchi*. Achetez Braun. (Publicité 1989)
(11) (a) On a chaud, et puis on sort, on respire *froid*, et on s'enrhume. (entendu en 1984)
(b) Les chevrettes aiment le soleil et respirent *pressé*. (corpus Grundt)

Pour étayer les oppositions entre les interprétations 'objet' (a) et 'manière' (b), l'auteur propose les tests suivants : dans le premier cas, une paraphrase en '*quelque chose de*

[5] Nous utiliserons ce terme traditionnel pour renvoyer indifféremment à tout dépendant verbal exprimant la manière, que son statut soit argumental (complément), ou non-argumental (adjoint/circonstant).

[6] Le concept de manière reste, ici comme chez de nombreux autres linguistes, assez vague et peut être défini *grosso modo* comme la caractérisation du procès.

Adj.' ou en 'du N_{Adj}' doit être possible (acheter *quelque chose de beau, du beau ;* respirer *du froid*) ; dans le second cas, la construction [V+Adj.Inv.] doit pouvoir être paraphrasée en 'avec N_{Adj}'/Adv.-*ment*[7], ou encore par la locution verbale 'avoir/faire Nv Adj.' (acheter *avec réflexion,* respirer *avec précipitation ; faire un achat réfléchi, avoir une respiration pressée*)[8].

Il n'est pas possible dans les limites de ce travail de discuter un à un les tests proposés ; nous observons cependant, à la suite de Noailly elle-même, que ces paraphrases fonctionnent « approximativement » (p. 108), et, comme elle, nous observons qu'elle met à l'écart les exemples qui ne sont pas clairement affiliables à l'une ou l'autre interprétation : « il est (des cas) moins clairs, voire d'indécidables » (p.109). C'est que, une fois encore, pour elle, l'interprétation ne peut qu'être « tantôt manière, tantôt objet ».

3 Limites aux tests de Noailly : le cas des emplois absolus de verbes transitifs

Nous nous interrogeons donc sur l'interprétation qu'il faudrait donner sur la base de l'exemple (12), extrait de notre corpus, si l'on applique le test en « faire » proposé par Noailly :

(12) Au supermarché, j'achète *utile*[9] 'je fais ***un achat*** *utile*'

Si l'on accepte cette paraphrase comme nous le faisons, il paraît difficile de décider à quoi réfère le nom 'achat' : *objet* acheté, ou *action* d'acheter ? Hors contexte, le décodage semble difficile, et les tests complémentaires proposés par Noailly ne semblent pas d'un très grand secours : **acheter avec utilité* est irrecevable, et on peut légitimement se demander si la paraphrase *acheter utilement* est le strict équivalent de (12) (cf. note 5).

Enfin, Noailly considère que, puisque la paraphrase en « faire/avoir » est possible aussi bien pour les verbes transitifs que pour les verbes intransitifs, on a affaire, dans le cas des verbes transitifs qui acceptent la paraphrase en « faire/avoir », à « une intransitivation » du verbe, qui est dit alors « en emploi absolu ».

Nous ne souscrivons qu'en partie à cette analyse ; lorsque Vtr est suivi d'un adjectif qui caractérise un objet non exprimé (*i.e.*, le test en *quelque chose de Adj./du* N_{ADJ} fonctionne), nous postulons une position 'zéro'[10] à droite de Vtr. Le rôle de l'adjectif est alors de sous-catégoriser l'entité/la classe d'entités à laquelle réfère $\varnothing_{OD}$. Dans ces cas, on a affaire à ce que M. Larjavaara (2000) désigne comme des « objets latents co(n)textuels » ou « extraco(n)textuels » : l'objet $\varnothing_{OD}$ est spécifique, identifiable en co(n)texte ou récupérable dans le savoir extralinguistique des locuteurs.

Là où nous nous éloignons de la position de Noailly (1994), c'est dans l'idée que pour l'emploi absolu il s'agisse d'un cas d'*intransitivation* : l'auteur explique qu'il n'y a alors plus aucun argument objet 'zéro' à droite du verbe ; « on ne pose pas précisément d'actant objet : l'énoncé en lui-même le présente comme totalement indifférencié » (p.112)[11]. Pour notre part, si nous sommes d'accord avec l'idée que l'actant objet est totalement indifférencié, nous estimons nécessaire de conserver une position $\varnothing_{OD}$ à droite de Vtr *quand il est suivi de Adj.Inv.*

C'est justement parce qu'elle n'observe les faits que sous le prisme de la transitivité verbale sans prendre en compte suffisamment le rôle de l'adjectif à droite du verbe que Noailly en arrive à la conclusion que [Adj.Inv.] *soit* caractérise $\varnothing_{OD}$, *soit* caractérise seulement le « sémième » particulier du verbe (la manière). Cette interprétation de la construction ne permet d'aucune manière d'analyser des énoncés comme (9) et (10) extraits de notre corpus :

(13) Julie cuisine *vietnamien.* (titre d'article dans un blog)
(14) Cuisinez *chinois* pas à pas. (titre de livre)

Vu que dans les deux cas, l'objet du verbe est totalement indéterminé, et non spécifique, il faudrait considérer selon son analyse que le verbe *cuisiner*, en emploi absolu dans les deux cas, n'a aucun souvenir de l'actant objet de sa structure argumentale de base. Il n'y aurait donc

[7] Accepter une telle paraphrase revient donc à dire qu'il y a stricte équivalence entre [Adj.Inv.] et Adv.-*ment*, ce qui est contradictoire dans l'analyse de Noailly, si on admet que les [Adj.Inv.] ne sont pas que de simples formes morphologiquement tronquées de l'[Adv.-*ment*] dans ces cas précis.

[8] Ces paraphrases sont reprises à l'auteur.

[9] http://www.santemagazine.fr/au-supermarche-jachete-utile-et-malin-29665.html

[10] Que nous notons $\varnothing_{OD}$ dans notre analyse.

[11] Plus loin, elle affirme qu'il s'agit « d'une réduction d'un actant » (p.113).

pas de position $\varnothing_{OD}$, et l'adjectif ne caractériserait que la *manière* de faire le procès.

Le problème est que les deux paraphrases proposées par Noailly fonctionnent aussi bien l'une que l'autre : on peut ***faire*** *de la cuisine vietnamienne* ou *chinoise*, mais on peut aussi cuisiner ***quelque chose*** *(des plats) de vietnamien ou de chinois*[12].

S'il est impossible de trancher aussi définitivement, c'est, nous semble-t-il, parce que la position $\varnothing_{OD}$ doit absolument être conservée dans le cadre de la construction [Vtr+Adj.Inv.] même s'il n'y a de référent accessible ni dans l'énoncé, ni dans l'univers de discours du locuteur[13]. Autrement dit, on a certes affaire là à une classe d'objets totalement indéterminés non spécifiques, mais ils restent bien présents à cause de la caractérisation adjectivale qui restreint cet ensemble auquel réfère la position « zéro » à droite de Vtr, en le sous-catégorisant ; en réalité, il faudrait même prendre le problème dans l'autre sens et partir de la définition lexicale du verbe, pour mieux comprendre ce qui se passe. En ce sens, le $\varnothing_{OD}$ attendu à droite du verbe *boire* (classe des liquides) ne peut pas être le même qu'à droite du verbe *manger* (classe des aliments, mais aussi des plats, etc.). Pourtant on trouve le même adjectif *bio* à droite de ces deux verbes en emploi absolu dans notre corpus, et l'adjectif ne réfère pas au même ensemble dans les deux cas. Cela signifie que malgré les emplois absolus de Vtr, un ensemble d'entités est *prévu* dans la structure argumentale du verbe, et par contrecoup, Adj. le caractérise. Il est donc nécessaire de conserver cette position $\varnothing_{OD}$ dans le cadre de la construction [V + Adj.Inv.], même si Vtr est en emploi absolu, ne serait-ce que pour mettre en évidence le rôle de l'adjectif à l'intérieur de la construction.

Consécutivement, il n'est plus nécessaire de maintenir la dichotomie entre emplois objet et emplois manière : dans les exemples (13) et (14) ci-dessus, la caractérisation adjectivale apportée par [Adj.Inv.] porte sur l'objet créé à l'issue du procès aussi bien que sur l'objet envisagé comme thème sémantique de l'action de *cuisiner* (des plats, qui peuvent être vietnamiens ou chinois, pour l'objet créé à l'issue du procès ou encore des aliments typiques des régions du monde concernées pour l'objet, thème sémantique du procès) et n'exclut absolument pas la caractérisation d'une certaine manière de réaliser le procès (on peut ainsi cuisiner (des aliments) *à la mode* chinoise ou vietnamienne). Quoiqu'il en soit, c'est [Adj.Inv.] qui impose le maintien de $\varnothing_{OD}$, et comme on peut le voir, il a la capacité de référer à des sous-ensembles d'entités marquées par le sceau de la stéréotypie : au seul niveau de sa définition lexicale, le verbe *cuisiner* n'implique en position d'objet syntaxique qu'une classe hyperonymique d'entités comestibles, et c'est par le jeu de la sous-catégorisation opérée par l'adjectif qu'il y a restriction à des sous-ensembles tels que les classes distinctes « aliments » ou « plats » ; c'est ainsi qu'on distinguera avec succès l'énoncé (14) de l'énoncé (15) :

(15) Cuisinez *chic* avec les grands chefs ! (Titre d'un livre)

Dans l'énoncé (15) [Adj.Inv.] renvoie à une *manière de* cuisiner (d'une manière chique), mais il opère *aussi* une sous-catégorisation de l'ensemble *plats* qui peut résulter du procès, ce qui semble impossible pour ce qui est de la classe des *aliments* (il n'existe *a priori* pas de sous-classe « **aliments** *chics* », comme le confirme notre consultation du web, alors qu'on peut plus facilement envisager une sous-classe de « **plats** *chics* », soigneusement préparés et présentés, comme le suggère l'intervention de grands chefs cuisiniers).

Comme on peut le voir, les phénomènes de portée sémantique de l'adjectif en jeu dans la construction [V+Adj.Inv.] sont très riches ; l'emploi absolu d'un verbe transitif ne doit pas être pris isolément du rôle que joue l'adjectif dans sa structure argumentale : si on consent à maintenir la position $\varnothing_{OD}$ à droite de Vtr, on peut voir les choses sous un autre angle que Noailly, ce qui ouvre des perspectives d'analyses encore plus riches du côté de l'incidence de l'adjectif sur ce $\varnothing_{OD}$, mais aussi du côté de l'expression de la manière.

4 Pour relancer l'analyse : une étude de cas

Nous achevons notre parcours en montrant que l'adjectif invarié à droite de Vtr peut à la fois

[12] Il est à noter une possible hésitation pour savoir quelle est l'entité hypéronymique $\varnothing_{OD}$ à la tête de la classe déclenchée par [Adj.Inv.] à droite de Vtr : on touche là à des phénomènes liés aux connaissances extralinguistiques du locuteur, que l'on retrouvera plus loin sous la notion de stéréotype.

[13] Ou encore dans ses connaissances extralinguistiques.

porter sur la manière et sur l'objet à partir de l'étude du cas du verbe transitif *cuisiner*, recensé dans notre corpus avec 21 adjectifs différents à sa droite, dont les principaux sont *japonais, bio, chic, sain, pratique, indien, solidaire, écolo*.

Nous rappelons d'abord la définition de la manière construite dans Stosic & Moline (2016) ainsi que la répartition des circonstants proposée par Melis (1983) pour ensuite proposer une analyse appliquée à [*cuisiner* + Adj.Inv.] sous forme de tableau synthétique.

Suite à leurs nombreuses recherches sur l'expression de la manière en français, Stosic et Moline (2016 : 184) définissent la manière en ces termes :

« La manière est une valeur sémantique complexe, incidente à un élément support, élaborée par des moyens lexicaux, syntaxiques, morphologiques, grammaticaux ou prosodiques et qui consiste en la diversification d'un procès, d'un état ou d'une qualité par une spécificité qualitative. »

L'application d'une « spécificité qualitative » au procès par l'utilisation de l'adjectif correspond précisément à l'expression de la manière. Cette qualité peut porter soit sur le procès directement (ce que Noailly avait précédemment identifié comme l'élément « sémiématique » du verbe) soit sur les circonstances liées au déroulement du procès. Or Melis (1983) propose une analyse très fine des relations possibles entre le sémantisme du verbe et les compléments de manière qui ne sont pas des actants du verbe (dits « circonstants ») et distingue :

- les compléments d'attitude :

(16) Marie roulait *avec anxiété* sur la N90 ;
(17) Pierre répondit *avec véhémence* à son détracteur.

- les compléments aspectuels

(18) Il a *rapidement* atteint son but.
(19) Il s'endort *progressivement*.

- les compléments instrumentaux :

(20) Il a peint le plafond *à la brosse*.
Nous observons que Melis intègre dans les instrumentaux les compléments construits avec des noms abstraits, généralement considérés comme compléments de moyen :
(21) Le soldat a défoncé la porte *d'un coup de pied.*

- les compléments sémiématiques[14], répartis en quatre catégories selon que la caractérisation du complément exprime :
- la qualité du procès :
(22) Madame écrit *élégamment*.

- un jugement évaluatif du procès :
(23) Il dessine *admirablement*.

- l'intensité :
(24) Il l'aime *éperdument*.

- la quantification du procès :
(25) *Il travaille énormément.*

En projetant cette grille sur les Adj.Inv. à droite du verbe transitif *cuisiner*, on obtient le tableau synthétique en annexe. Plusieurs conclusions peuvent être tirées de ce tableau :

(i) [Adj.Inv.] porte sur l'objet $\varnothing_{OD}$ présent sous forme de thème sémantique (*bio*, *casher*), sous forme de résultat de l'action (*végétarien*, *végétalien*), ou encore sur les deux en même temps (*japonais* et les autres Adj.Rel. de nationalité, *sain* et *léger*). On peut opposer ce fonctionnement aux cas où l'adjectif ne caractérise pas $\varnothing_{OD}$ (*chic*, *pratique*, *solidaire*, *écolo*, *vert*, et *durable*).

(ii) Aucun des adjectifs ne modifie l'aspect, ni n'exprime une quantification, une valeur d'intensité ou un jugement sur le procès verbal, à droite du verbe *cuisiner*.

(iii) Dans tous les cas, lorsque [Adj.Inv.] caractérise $\varnothing_{OD}$, on constate qu'il spécifie la manière (le complément instrumental, le plus souvent) : *japonais*, *bio*, *sain*, *casher*, *végétarien*, *végétalien*, *léger* sont des propriétés de l'objet $\varnothing_{OD}$ subissant l'action exprimée par le verbe, ou de son résultat ; ce sont en même temps des propriétés des

[14] Le terme « sémiématique » s'inspire de l'opposition entre « taxième » et « sémième » de Damourette & Pichon (1911-1940), celui-là renvoyant en gros, et sans que la coupure soit radicale, au matériau grammatical, celui-ci au matériau lexical.

ingrédients qui servent à réaliser l'action (des sauces, ou des ingrédients qui entrent dans la composition des plats, par exemple) ; il est à noter ici encore que le phénomène de la stéréotypie entre en jeu : certains ingrédients sont typiques de la cuisine japonaise, indienne, etc.

5 Conclusion

Nous avons donc montré que non seulement l'adjectif en emploi adverbial à droite du verbe transitif ne modifie pas exclusivement la manière de réaliser le procès OU un objet « latent » présent sous sa forme $\varnothing_{OD}$, mais qu'il a la capacité de modifier les deux en même temps, pour peu qu'on consente à maintenir une position $\varnothing_{OD}$ à droite de V, même lorsqu'il est en emploi absolu.

Ce fonctionnement atypique de l'adjectif à droite d'un verbe n'est cependant pas limité aux seuls verbes transitifs et nos prochains travaux auront pour objectif de voir s'il existe des points communs entre la spécification de la manière par un adjectif invarié à droite d'un verbe transitif et à droite d'un verbe intransitif comme *voyager* (ex. *voyager malin, léger, responsable, bio, chic…)*.

Nous aurons par ailleurs à rendre compte des spécificités de l'adjectif en emploi adverbial par rapport à l'adverbe en *–ment*. Si dans certains cas l'adjectif semble remplir des lacunes lexicales (ex. *voyager *responsablement*), dans d'autres il a plutôt pour rôle d'exprimer une valeur différente de celle véhiculée par l'adverbe *(penser chinois/chinoisement)*. Les deux cas de figure témoignent d'importantes particularités syntactico-sémantiques de ce type d'emploi de l'adjectif.

Références

Abeillé, A. & Godard, D. (2004), « Les adjectifs invariables comme compléments légers en français », in *L'adjectif en français et à travers les langues*, Caen, PUC, pp.209-224.

Damourette, J. & E. Pichon (1911-1940), *Des mots à la pensée. Essai de Grammaire de la Langue Française.* Paris, d'Artrey.

Grundt, L.O. (1972), *Études sur l'adjectif invarié en français*. Bergen-Oslo, Universitets-Forlaget.

Guimier, Cl. (1996), *Les adverbes du français : le cas des adverbes en –ment*. Paris/Gap, Ophrys, collection « L'essentiel français ».

Guimier, Cl. & Oueslati, L. (2006), « Le Degré de figement des constructions 'Verbe + Adjectif Invarié' », in *Composition syntaxique et figement lexical*, Presses Universitaires de Caen, pp.17-37

Larjavaara, M. (2000), *Présence ou absence de l'objet, Limites du possible en français contemporain*. Thèse pour le doctorat présentée à la Faculté des Lettres de l'Université de Helsinki, Université de Helsinki.

Melis, L. (1983), *Les circonstants et la phrase : étude sur la classification et la systématique des compléments circonstanciels en français moderne*. Louvain, Presses Universitaires de Louvain.

Moignet, G. (1962), « L'incidence de l'adverbe et l'adverbialisation des adjectifs ». *Travaux de Linguistique et de Littérature* 1. Strasbourg, Université de Strasbourg.

Moline, E. & Stosic, D., (2016), *L'expression de la manière en français*. Paris, Ophrys, collection « L'Essentiel français ».

Noailly, M. (1994), « Adjectif adverbal et transitivité », in *Cahiers de grammaire*, n°19 (pp.103-114)

Riegel, M., Pellat, J.-C., & Rioul, R. (1994), *Grammaire Méthodique du Français*. Paris, P.U.F.

Annexe

Tableau 1 *Les adjectifs entrant dans la construction 'cuisiner + Adj.Inv.'*

Manière							**Objet**	
Cpt Att	Cpt Asp.	Cpt Inst./Moy.	Compléments sémiématiques				$\varnothing_{OD}$ thème (aliment transformé)	$\varnothing_{OD}$ Résultat (plat obtenu)
			Qual.	Quant	Int.	Éval.		
-	-	*japonais*	*japonais*	-	-	-	*japonais*[15] (aliments stéréotypiques)	*japonais* (stéréot. de plat)
-	-	*bio*	-	-	-	-	*bio*	-
(*chic* ?)	-	*chic*	-	-	-	-	-	-
-	-	*sain*	*sain*	-	-	-	*sain*	(*sain*)
-	-	*pratique*	*pratique*	-	-	-	-	-
(*solidaire* ?)	-	-	*solidaire*	-	-	-	-	-
-	-	*écolo*	*écolo*	-	-	-	-	-
-	-	*vert*	-	-	-	-	-	-
-	-	*durable*	-	-	-	-	-	-
-	-	*casher*	-	-	-	-	*casher*	-
-	-	*végétarien*	*végétarien*	-	-	-	-	*végétarien*
-	-	*végétalien*	*végétalien*	-	-	-	-	*végétalien*
-	-	*léger*	-	-	-	-	*léger*	*léger*

[15] Dans notre corpus, sont trouvés, et analysés de la même manière : *chinois*, *indien*, *italien*, *vietnamien*, ou encore *oriental*

Morphological Ambiguities in Egyptian Arabic Dialect Used in Social Media

Reham Marzouk
Phonetics and Linguistics Dep.,
Faculty of Arts, Alexandria University
P.O BOX 21526, Alexandria, Egypt
marzoukreham@gmail.com

Seham El Kareh
Phonetics and Linguistics Dep.,
Faculty of Arts, Alexandria University
P.O BOX 21526, Alexandria, Egypt
sehamelkareh@gmail.com

Résumé/Abstract

This study aims to reveal the main morphological ambiguities occurs during the morphological analysis of the Egyptian Arabic Dialect (EGY) in particular its written form used in social media and how far of morphological analyzers are able to handle such ambiguities. Thus it evaluates the automatic annotation of the Egyptian Arabic Penn-Treebank ARZ ATB which are collected by Linguistic Data Consortium LDC and analyzed using the Columbian Arabic diaLectal Morphological Analyzer CALIMA. The results showed that several ambiguities couldn't be handled during the morphological analysis. Moreover, the error analysis proved that the major reason of morphological ambiguity of the Egyptian Arabic dialect is the Orthographic variations of its written form. These variations reflected the lack of an authorized writing system governs the written form of the dialect.

1 Introduction

Arabic language is known as one of the Semitic language family (Holes, 2004), which is used by more than 300 millions native speakers (Dasigi & Diab, 2011), (Retso, 2013). The prominence of the Arabic language is the existence of several varieties of the language that are used for different purposes. Modern standard Arabic (MSA) is the modern descendant of Classical Arabic (CLA), the language of the Islamic holy book (Holes, 2004). MSA, nowadays, is used in all the writings all over the Arab world, and its spoken form dominates all the media, in addition to learning it at schools. On the other side, spoken Arabic dialects represent the Arabic language varieties that are used in the daily communication activities (El-Hassan, 1977). Each Arabic country has its own dialect that is labeled by Badawi (1985) as (educated spoken), (Ibrahim, 2009).

Nowadays, spoken dialects are intervened, and used in a wide range of written texts due to the spread of the social medial channels such as SMS, chatting, and other communication mediums which became rich resources for these dialects in its written form (Dasigi & Diab, 2011).

Accordingly, processing these dialects became imperative to develop applications such as morphological analysis, classification, machine translation,...etc.

This work emphasizes on the influence of the social media usage on the Arabic Language, as well as, its dialects. It is a profound morphological study of the electronic texts written by Egyptian Arabic dialect, in specific, aims to clarify the causes of morphological ambiguities that accompanied the existence of such electronic texts. Hence, the study is considered as a preliminary step to provide methods for further handling such morphological ambiguities. Therefore, ARZ ATB Penn Treebank Corpus is used in this research to represent the Egyptian Arabic dialect which is considered as the most prevalent dialect used in electronic texting among the Arab world. ARZ ATB corpus is gathered by LDC, University of Pennsylvania. Then, it is morphologically annotated using the morphological analyzer CALIMA.

In this paper, the morphological analysis of ARZ ATB corpus is evaluated and errors are

classified to investigate the analyzer's proficiency in analyzing the written form of the Egyptian Arabic as it appears in social media channels such as SMS, discussion forums, Whatsapp, etc. The main contribution of this research is that the source of the texts is different from the usual texts used to present the written forms. Furthermore, the results reveal a requirement of modified methods to handle the morphological ambiguities. Thus, this study is undertaken as a first stage in implementing a system to handle such ambiguities.

The research is organized as follows: section 2 overviews the main related works that conducted for the morphological analysis and disambiguation of the Arabic dialects. Section 3 describes briefly the significant features of the Egyptian Arabic morphology Section 4 explains the the role of social media texts in natural language processing. Section 5 introduces the procedure of analyzing the corpus. Finally sections 6 and 7 display the results and the conclusion of the study.

2 Related Studies

In the last decades, several morphological analyzers for Arabic language were developed based on different approaches. Most applications are applied on MSA. However, some morphological analyzers were developed in order to handle the different Arabic dialects such as Levantine Arabic and Egyptian Arabic. Some of these morphological analyzers were evaluated by their developers and others evaluated by others associations. Habash, (2009) built MADA+ TOKEN that includes part-of-speech tagging, diacritization, lemmatization, disambiguation, stemming, and glossing. It consists of two components: MADA that adds lexical and morphological information and TOKAN that generates a tokenization to tokenize the words and identify its stem. MADA has over 96% accuracy on morphological analysis and lemmatization, and over 86% accuracy in predicting full diacritization. Arfath Pasha et al., (2014), also presented the morphological analyzer MADAMIRA, the system combines the best aspects of previous two systems: MADA+AMIRA, it has the same general design of MADA with additional components inspired by AMIRA. MADAMIRA is designed to analyze MSA and EGY. The accuracy of the system was 80% for MSA and 76.4% for EGY (Pasha, et al., 2014).

3 Egyptian Arabic Morphology

Arabic language differs in terms of the typography from Latin, it is comprised of 60 characters including letters, diacritics, punctuation marks (Attia, 2008). Diacritic marks that refer to short vowels were omitted from MSA and Arabic dialects written texts, whereas long vowels are only written using the 3 sounds {*A, iy, uw*}. Moreover, Arabic is a highly inflectional language with complicated morphological system (Attia, 2008). EGY has the same morphological aspect with slight changes. For instance, the deletion of the case ending that refers to different cases of the word: nominative, accusative, and genitive (Gadallah, 2000).

The main difference between EGY and MSA is the dialectal vowel system (Holes, 2004). While some MSA words were preserved in EGY, others have undergone phonological changes such as: long vowel shortening, deletion of final glottal stop />/, and Monophthongization (turning diphthongs of MSA into one long vowel) (Gadallah, 2000).

Arabic Language is a clitic language. Clitics are morphemes that have the syntactic characteristics of a word, but are bound to other words. The perfect examples for clitics are conjunctions, prepositions and particles, and pronouns that could be attached to the word either at its beginning or its end (Attia, 2008).

Definite article in EGY is the prefix /l/. Since Egyptian Arabic doesn't allow consonant clusters in the onset, /i/ is inserted and a glottal stop is epenthesized. When the definite article is preceded with preposition, the epenthetic glottal stop is deleted (Watson, 2000).

Gender and number in EGY are defined by suffixes to refer to feminine singular feminine, masculine dual, feminine dual, masculine plural and feminine plural.

Broken plural is another sort of plural that is constructed by 'changing the shape of the singular through various morphological process such as long vowel insertion, consonant gemination, semivowel insertion and the affixation of consonant additional to those of the root' (Holes, 2004).

4 Processing Social Media Text

The growing popularity of social media produced enormous quantities of daily electronic texts. These texts act as data for many applications such as information extractions, linking, classification, POS tagging, etc. (Habib, 2014). The Annual Arab social media survey (2015), produced by the Dubai school of governance and innovation, reported that Egyptians are one of the highest users of social media with (94%). Egyptian Arabic used in social media is much different from other written genres, since 'its vocabulary is informal with intentional deviations from standard orthography such as repeated letters for emphasis; typos and non-standard abbreviations are common; and non-linguistic content, such as laughter, sound representations, and emoticons' (Bies et al., 2014).

Board Operational Language Translation program (BOLT) is produces by DARPA, Defense Advance Research Project Agency, and intends to develop technology to translate information from informal foreign language sources. A stage of achieving this project was developing an annotated Egyptian Arabic TreeBank (ARZ ATB). The corpus is collected by LDC from different social media channels (Maamouri et al., 2014).

Thereafter, the annotation of ARZ Penn TreeBank went through POS/morphological annotation. CALIMA is the Egyptian Arabic morphological analyzer which was used for the automatic annotation (Maamouri et al., 2014). CALIMA refers to the Columbia Arabic Language and dIalect Morphological Analyzer. The system is built by extending the Egyptian Colloquial Arabic Lexicon (ECAL) (Habash et al., 2012). It consists of six tables, three tables specify the complex prefix/suffix and stems. And three tables specify compatibility across the class categories (prefix-stem, prefix-suffix and stem-suffix), figure 1, (Habash et al., 2012). The annotation by CALIMA follow the LDC POS guidelines and the Conventional Orthography for Dialect Arabic CODA (Habash et al., 2012).

wl	wali	NPref-Li	and + for/to <pos>wa/CONJ+li/PREP+</pos>
ll	lil	NPref-Li	to/for + the <pos>li/PREP+Al/DET+</pos>
wll	walil	NPref-Lil	and + to/for + the <pos>wa/CONJ+li/PREP+Al/DET+</pos>
wbAl	wabiAl	NPref-BiAl	and + with/by the <pos>wa/CONJ+bi/PREP+Al/DET+</pos>

Fig 1- The compatibility table

5 The Analysis

To evaluate the annotation of ARZ ATB, we created a gold standard to be compared with CALIMA's results (Sawalha, 2011). The gold standard was built by using the most frequent 6543 word types selected randomly from the ARZ ATB corpus, and it is produced in the same format of CALIMA's output. Words are inserted in separated lines, including their detailed morphosyntactic information, such as: the lemma and the vocalization of the word. The morphological information of each word in the gold standard was provided manually to present the model analysis of it, figure 2.

Word Type	Freq	Vocalization	Lemma	Proclitic	Prefix	POS	Suffix	Enclitics
الله	131	Al~ah	Al~ah	0	DET	NOUN-PROP	0	0
ربنا	116	rab~inA	rab~	0	0	NOUN	0	PRON_1P
حاجة	103	Hajap	Hajap	0	0	NOUN	NSUFF_SG	0
كل	90	kul~	kul~	0	0	NOUN_QUANT	0	0
مع	79	maEa	MaEa	0	0	NOUN	0	0
كده	70	Kidah	Kidah	0	0	NOUN	0	0
حد	55	Had~	Had~	0	0	NOUN	0	0
كدا	55	Kidah	Kidah	0	0	NOUN	0	0
ماما	54	mAmA	MAmA	0	0	NOUN	0	0
ممكن	55	Mumkin	mumkin	0	0	ADJ	0	0
جدا	47	jid~AF	jid~	0	0	NOUN	0	CASE_ACC_INDEF
غير	46	Giyr	Giyr	0	0	NOUN	0	0
كتير	45	kitiyr	Kitiyr	0	0	ADJ	0	0
طيب	44	Tayib	Tayib	0	0	ADJ	0	0
والله	44	wal~ahi	Al~ah	CONJ	DET	NOUN_PROP	0	0
بعد	43	baEd	baEd	0	0	NOUN	0	0
محمد	43	maHamad	maHamad	0	0	NOUN_PROP	0	0
الناس	41	Ain~As	Nas	0	DET	NOUN	0	0
كمان	39	Kaman	kaman	0	0	NOUN	0	0
واحد	36	waHid	waHid	0	0	NOUN_QUANT	0	0
كويس	35	Kuwayis	kuwayis	0	0	ADJ	0	0
تاني	32	tAniy	vAniy	0	0	ADJ	0	0
عمرو	32	>amr	>amr	0	0	NOUN_PROP	0	0
قبل	32	Qabl	Qabl	0	0	NOUN	0	0

Fig 2-The gold standard

The selected words of gold standard was annotated again using CALIMA and the results were sorted in different lists, each one presents a certain morphological features. Each morphological feature was evaluated separately to provide more detailed results.

Word	LEMMA1	LEMMA2	LEMMA3	LEMMA4
مثالية	mivAliy	0	0	0
محمود	maHmuwd	maHmuwd	0	0
محضر	miHaD~ar	maHDar	HaD~ar	HaD~ar
محبس	maHbas	Habas	0	0
مرات	Marap	0	0	0
محظوظ	maHZuwZ	maHZuwZ	0	0
مستخبي	mistaxab~iy	0	0	0
مريض	mariD	ray~aD	0	0
مسكن	musak~in	0	0	0
مشغولة	maSguwl	0	0	0
مصورين	muSaw~ir	miSaw~ar	0	0
مصححة	miSaHSaH	0	0	0
معروف	maEruwf	maEruwf	0	0
معلش	maEliS	0	0	0

Word	POS 1	POS 2	POS 3	POS 4	POS 5	POS 6
تاوي	ADJ	NOUN	0	0	0	0
نافع	ADJ	NOUN	0	0	0	0
نادي	NOUN	CV	PV	NOUN_PROP	IV	ADJ
نفسيا	ADJ	0	0	0	0	0
هادي	ADJ	PV	IV	NOUN	0	0
كلية	NOUN	PV	0	0	0	0
هدايا	NOUN	0	0	0	0	0
هاديه	ADJ	NOUN	IV	0	0	0

Fig 3 & 4- Samples of CALIMA results

The morphological analysis of ARZ ATB was composed of five essential processes: vocalization, normalization, lemmatization, tokenization, and POS tagging. The output of each process was separately examined to obtain detailed explanations for the origins of such errors that are caused due to the system's shortage to cover all the distinctive characteristics of the written Egyptian Arabic form. These errors occasionally concluded morphological ambiguous analyses for the same word. Therefore, the observed errors that occurred during each process were listed and discussed separately as follows

5.1 Vocalization (Diacritization)

Vocalization is the process where suitable diacritics are interpolated to the undiacritized words. Wrong diacritization have been observed during the analysis due to the inability of the system to cover all the phonological alternation rules that dominate the language under certain conditions such as:

- Deletion of the epenthetic glottal stop of the definite article when it is preceded by preposition, since preposition in EGY are open classes ending with vowels, e.g. {*fi+Al+bayt*} becomes {*filbayt*}, (in the house).

- Assimilation of the definite article in case of being followed by coronal consonants, ex: {*Al+nAs*} becomes {*An~As*}, (the people).

- Regional Dialects, due to the great similarity among these dialects, are interfered during the annotation. For instance, سمع {*simiE*}, (to hear) in the Cairene, and سمع {*samaE*} in the Alexandriane.

5.2 Tokenization(segmentation)

Tokenization in Arabic language requires to segment the joined affixes in the word. Hence, the causes of ambiguities that affect the accuracy of the tokenization were categorized as following:

- Spelling variance: due to the inconsistency of Egyptian Arabic written form. For instance, the omission of the definite article /l/, after a preposition by some writers as: بكتاب {*bikitAb*} instead of بالكتاب {*bilkitAb*}, (with the book).

- Homography between word after their attachment with certain morphemes, e.g. the noun بكره {*bukrah*} and the verb بكره{*bi+>a+krah*}

- Overgeneralization: sometimes parts of the words are tokenized wrongly leaving undesired tokens with no sense, e.g. the noun بشره {*ba$r+ap*} (skin) can be segmented into بشره {*bi$ar~+ap*}

5.3 Orthographic Lemma Identification

CALIMA shows a high accuracy in identifying the lemma of the tested words except in some cases of broken plurals that are hardly lemmatized due to the lack of coverage of their different forms, e.g صحاب *{SuHAb}* (friends), and اخوان *{AixwAn}* (brothers).

5.4 POS Tagging

Part Of Speech (POS) tagging covers the parts the Egyptian Arabic word:

[proclitic1][proclitic2][prefix]
[stem][suffix][enclitic1]enclitic2]

The in-existence of standard writing system for the Egyptian dialect led to many replacements among their consonants, vowels and morphemes. Thus, the system confronted a challenge in identifying parts of the words due to the resemblance between some of these replaced morphemes. Consequently, that produced several uncertain tags for these morphemes. Some instances for the replacements are:

- Using the same grapheme to write the consonant أ *{>a}* and the long vowel ا *{A}*. Therefore, some words are confused with others e.g. the adj بارد *{bArid}* and

verb بارد *{ba+>arud}* which are written identically.

- Using the same grapheme to write the distinct two consonants ة {*ap*} and ه {*h*}, which led to confusion between the singular feminine suffix {*ap*} and the possessive pronoun enclitic {*uh*}, e.g. كتابه {*kitAp+uh*} (his book) and كتابه {*kitAb+ap*} (writing).

- Shortening long vowels due to the phonological alternation rules that govern the Egyptian dialect and have been transferred to the written form of the dialect. This alternation also caused many ambiguous words such as the Adj سمعه {*samEah}*, (hearing) and سمعه {*samaEuh}*, (he heard him).

- Replacing emphatic consonants with non ephatic consonants, e.g طرابيزة {*Tarabizap*} and ترابيزة {*tarabizap*} (a drum).

6 Results

Recall, precision, accuracy and F-score were measured for the output of each category in the tested data. a normalization from 0 to 1 were achieved for words with more than one analysis and the results were summarized in the table 1.

Errors, resulted due to ambiguous analyses were classified to clarify the major causes of these ambiguities. Hence, our classification revealed that 40.4% of the errors are attributed to the orthographic variations and that is the highest percentage, whereas the remaining errors were caused due to other reasons such as wrong tags, lack of broken plural coverage and typography.

TABLE 1- Measurement Results

Feature	Recall	Precision	F-score
POS	83%	82.5%	82.5%
Inflection	99.5%	82.5%	90.1%
Definiteness	99.4%	94.2%	97%
Proclitics	99.1%	71.6%	83.3%
Enclitics	99.1%	93.7%	93.7%

7 Conclusion and Future Work

Classifying the reasons beyond the ambiguities that may rise during the morphological analysis process is considered as a step toward rendering specific solutions to handle these ambiguities. The conspicuous ambiguities in this stage was correlated This study attempts to provide a valuable resource for improving Egyptian Morphological analyzers through to the inconsistency of the Egyptian Arabic written form, because of the inclination of the writers to improvise, as well as, the lack of a specific writing system to rule the spoken dialects that are used in writing texts. Therefore, some writers follow the writing system of MSA, and others apply the phonological alternations of the spoken dialect on the written form. This inconsistency led to different variances for the same word. Some of them were considered as typography and others as orthographic variations due to their frequent occurring. These variations required a serious normalization process to map them into one standard form that match the lexical data of the morphological analyzer. This pre-processing stage is essential to diminish the unwanted analyses during the annotation of these sorts of texts. Thus, developing a normalization system, based on rewrite rules that map the occurring variations into a standard form, is our concern in the future.

References

Abdel-Massih, E. T., Abdel-Malek, Z. N., & Badawi, E. S. M. (1981). *A reference grammar of Egyptian Arabic*. Center for Near Eastern and North African Studies, Univ. of Michigan.

Ali, A., Mubarak, H., & Vogel, S. (2014). Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr. In *International Workshop on Spoken Language Translation (IWSLT 2014)*.

Attia, M. A. (2008). *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation* (Doctoral dissertation, University of Manchester).

Badawi, E. S., Carter, M., & Gully, A. (2013). *Modern written Arabic: A comprehensive grammar*. Routledge.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., ... & Rambow, O. (2014, October). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)* (pp. 93-103).

Dasigi, P., & Diab, M. T. (2011). CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic. In *IJCNLP* (pp. 318-326).

Eisenstein, J. (2013, June). What to do about bad language on the internet. In *HLT-NAACL* (pp. 359-369).

El-Hassan, S. A. (1977). Educated Spoken Arabic in Egypt and the Levant: A critical review of diglossia and related concepts. *Archivum Linguisticum Leeds*, *8*(2), 112-132.

Faaß, G., Heid, U., & Schmid, H. (2010, May). Design and Application of a Gold Standard for Morphological Analysis: SMOR as an Example of Morphological Evaluation. In *LREC*.

Gadalla, H. A. (2000). *Comparative Morphology of Standard and Egyptian Arabic* (Vol. 5). Munich: Lincom Europa.

Habash, N., Rambow, O., & Roth, R. (2009, April). MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*(Vol. 41, p. 62).

Habash, N., Diab, M. T., & Rambow, O. (2012). Conventional Orthography for Dialectal Arabic. In *LREC* (pp. 711-718).

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.

Habash, N., Eskander, R., & Hawwari, A. (2012, June). A morphological analyzer for Egyptian Arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology* (pp. 1-9). Association for Computational Linguistics.

Habib, M. B., & Van Keulen, M. (2014). Information extraction for social media. Association for Computational Linguistics.

Hassig, H. L. (2011). *Deriving Cairene Arabic from Modern Standard Arabic: A framework for using Modern Standard Arabic text to synthesize Cairene Arabic speech from phonetic transcription* (Master's thesis).

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Ibrahim, Z. (2009). *Beyond lexical variation in modern standard Arabic: Egypt, Lebanon and Morocco*. Cambridge Scholars Publishing.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., & Eskander, R. (2014). Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *LREC* (pp. 2348-2354).

Marzouk, R., (2016). *Disambiguating Egyptian Arabic Morphological Analysis: A Linguistc Study* (Master's thesis).

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014, May). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *LREC* (Vol. 14, pp. 1094-1101).

Resto, J., (2013). "What is Arabic," OWENS (ed), pp. 433-450, 2013

Salib, M. B. (1981). *Spoken Arabic of Cairo*. American University in Cairo Press.

Sawalha, M. S. S. (2011). *Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora*. University of Leeds.

Watson, J. C. E. (2007). The Phonology and Morphology of Arabic. The phonology of the world's languages, ed. J. Durand.

Le développement de l'organisation syntaxique et discursive en français L2 dans les productions orales des apprenants japonais : débutants aux avancés

Chieko KAWAI
Laboratoire FoReLL,
Université de Poitiers
ckawai33@gmail.com

Résumé/Abstract

De nombreuses recherches sur l'acquisition du FLE ont contribué à l'éclaircissement du développement des constructions grammaticales : la structure simple, caractérisée par la juxtaposition ou la coordination, se développe vers la structure complexe comme la subordination. Dans ce présent travail, basé sur l'observation des productions orales d'apprenants japonais adultes du français L2, je me pose la question de savoir si l'appropriation des constructions grammaticales de la langue cible chez les apprenants japonais s'effectue de la même manière que celle observée dans les études antérieures. Pour répondre à cette question, je me propose de montrer comment les apprenants, notamment débutants, organisent des énoncés complexes. Je vais observer également l'emploi de la structure de focalisation du type *c'est/il y a…qui/que* en rapport avec la construction relative dépourvue d'éléments introducteurs, et illustrer les types de structures se développant selon différents stades de l'acquisition et quelles sont les interlangues des apprenants japonais.

Mots-clés : Processus du développement syntaxique, Enoncés complexes, Interlangues, Productions orales, Différents stades de l'acquisition du FLE

1 Introduction

De nombreuses études sur l'acquisition du FLE (entre autres, Bartning 1997 ; Bartning et Schlyter 2004 ; Klein et Perdue 1997 ; Perdue 1984 ; Véronique 2009) ont démontré que la progression syntaxique s'observe premièrement dans une structure simple comme des juxtapositions dotées ou non des connecteurs comme *et*, *mais*, etc. et secondement dans une structure complexe qui se caractérise par la subordination[1]. La plupart des travaux portent leur analyse sur les apprenants anglophones, germanophones ou suédophones. En revanche, les études sur les apprenants japonais du FLE sont encore en nombre limité[2]. Notre objectif consiste donc d'une part à examiner comment se développe l'organisation syntaxique et discursive chez les apprenants japonais selon différents stades d'acquisition, et d'autre part à confirmer ou infirmer si le proces-

[1] D'après les recherches portant sur la CAF (Complexity, Accuracy and Fluency) dans l'acquisition de L2 (Housen et Kuiken 2009), le terme de la complexité est initialement employé pour renvoyer aux propriétés de tâche de langue (complexité de tâche) et aux propriétés de la performance et la compétence dans l'usage de L2 (complexité L2). La complexité L2 est à son tour envisagée dans deux perspectives : la complexité cognitive qui se caractérise par les difficultés relatives à la performance et l'acquisition de L2 se trouvant à l'échelle individuelle et la complexité linguistique qui reflète le rapport entre les caractéristiques de la langue et la performance/compétence de l'apprenant. Notre présente étude est plutôt basée sur la complexité L2.

[2] Pour les études récentes sur l'acquisition du FLE chez les apprenants japonais, il y a celles de Trévisiol(-Okamura) (2003 ; 2015 entre autres) et Granget (2014).

sus du développement syntaxique se montre identique à (ou proche de) celui qui a été observé dans les études antérieures sur les allophones des langues maternelles (LM), très éloignées de la LM des apprenants japonais. Pour ce faire, nous allons d'abord montrer comment les apprenants, notamment débutants, construisent des énoncés complexes. Nous observerons par la suite les structures de focalisation, qui ont un statut particulier et qui sont observées dès le stade débutant, en rapport avec l'emploi de relative simple. Enfin, nous nous concentrerons sur l'emploi de différents types de subordonnées qui semble montrer l'itinéraire développemental de la structuration des énoncés.

2 Analyse

2.1 Recueil des données et Apprenants

Les données orales de notre étude ont été recueillies auprès de 48 apprenants japonais, résidant au Japon ou en France, à partir d'un dialogue (question-réponse)[3] effectué en français avec une enquêtrice japonaise. Nous avons également demandé à 15 apprenants une production en monologue d'un récit fictif (*Histoire de Cendrillon*) : dans ce type de récit qui ne permet pas l'étayage de l'interlocutrice, les apprenants sont tenus de produire un énoncé fini. La sociobiographie des apprenants, comme la durée de l'apprentissage de la LC, etc., est hétérogène, de sorte que nous les avons classés selon quatre stades en nous référant à l'étude de Bartning et Schlyter (2004). Pour le récit personnel, les apprenants sont répartis en proportions équivalentes dans chaque stade, si on regroupe les deux stades avancés. En revanche, pour le récit de fiction, la plupart des participants se concentrent sur le stade intermédiaire et sur les deux stades avancés en raison de la difficulté à produire un récit pour les apprenants se situant au stade post-initial.

Tableau 1 : Stades de l'acquisition et nombre d'apprenants

Stade de l'acquisition	Post-initial (PI)	Inter-médiaire (IM)	Avancé bas (AB)	Avancé moyen	Total
Récit personnel	17 pers.	15 pers.	7 pers.	9 pers.	48 pers.
Récit de fiction	2 pers.	5 pers.	3 pers.	5 pers.	15 pers.

2.2 Illustration

Les apprenants débutants dans notre corpus tendent à employer la coordination en se servant des marqueurs grammaticaux, comme cela été signalé par les travaux antérieurs. Nous constatons toutefois des structures qui semblent en train de se développer vers la construction complexe, comme le montrent les énoncés suivants :

[3] Le corpus oral représente 546,64 minutes d'enregistrements et 53306 mots. En ce qui concerne la transcription, nous avons adopté les conventions, employées par l'équipe du GARS (*Groupe Aixois de Recherches en Syntaxe*), fondée par Blanche-Benveniste, Deulofeu, Jeanjean, Stéfanini et Valli, et l'équipe DELIC (*Description Linguistique Informatisée sur Corpus*).

(1) (YSH/PI/F/dia.): à la futur + hm ++ je ne je ne [s]- + je ne je ne XXX + oui donc je + ma ++ ah + [dəv]- je n'ai pas + décidé comme [dəvniʁ] + mais + *je voudrais* ah ++ hm ++ **c'est** faire le peinture + XXX longtemps + oui

(2) (MIF/PI/F/dia.) : hm ++ je + ce que je + ce que je suis heureuse ah ++ le gens de + mon foyer + eh + souvient mon nom + et appelle(nt) et moi ah *P* comme ça + oui c'est très + heureuse + oui

(3) (YOK/PI/F/dia.): hm ++ au + je ne sais pas à Paris mais + à Poitiers + eh c'est pas joli + parce que (rire) il y a beau- + *tomb/E/* **il y a** *beaucoup de + cacas* (rire) + oui dans le ah + sur le rue + donc (rire) + oui + pas bien (rire)

Les apprenants, dépourvus de moyens syntaxiques pour former un énoncé complexe, s'expriment avec les moyens lexicaux (1) ou syntaxiques avec la thématisation (2 et 3). A ce stade, les formes non analysées *c'est* et *il y a* dont l'emploi est très fréquent sont souvent utilisées pour pallier des problèmes syntaxiques (les parties notées en gras). Bien que les éléments grammaticaux *qui/ que* se manifestent chez certains apprenants de ce stade, ils ne semblent pas encore être assimilés :

(4) (FUK/PI/F/dia.) : maintenant j'habite + de + bâtiment de + ah + étudiant + mais c'est + rez-de-chaussée + donc ++ une fois + ah + quel- quelqu'un ++ **qui** + je ah **que** + je sais pas + ah ++ **qui** vient + ma chambre + et + i- il [di] [ke] + eh + le fenêtre + don- donc + j'ai peur + un peu +

L'emploi de la relative peut être encore instable dans les stades même au-delà du PI. Mais cette instabilité ne tient pas à la méconnaissance de la fonction des relatives et semble plutôt découler du problème d'organisation discursive[4] :

(5) (KEI/IM/F/dia.) : ah parce que + **j'ai** une amie qui est + qui est + qui est + (rire) + qui ++ a ++ la mère + qui est française + et ++ *elle* m'a raconté beaucoup + de ++ la vie français +

Dans cet énoncé, la relative s'enchaîne à partir de la formule « j'ai…qui » : *j'ai une amie [qui a la mère] [qui est française]*. Etant donné que le pronom sujet *elle* réfère à « la mère d'une amie », il ne s'agit pas d'apposition et chaque pronom *qui* renvoie à l'élément antéposé. A ce stade où d'autres pronoms relatifs, comme *dont*, ne semblent pas encore assimilés[5], le relatif *qui* fonctionne comme un marqueur commun qui sert à caractériser son antécédent.

- **L'emploi de *c'est/ il y a…qui/ que* et de la construction relative**

La structure introduite par les formes non analysées associées avec *qui/ que* s'observe dès le stade PI comme nous l'avons observé plus haut, et augmente de plus en plus avec l'avancement dans les stades. Observons le tableau suivant :

Tableau 2 : Répartition de constructions relatives dans notre corpus

Récit personnel :

	Constructions relatives avec les éléments introducteurs				Relatives sim-	Nombre total de mots selon les stades[6]
	il y a	c'est	j'ai	Pseudo-clivé	Autres contextes	
PI	4	3	1		6	17684
IM	6	8	3	3	18	15539
AB	**14**	**8**	1	5	19	9822
AM	**11**	**12**		3	32	10132
Total	35	31	5	11	75	53177

[4] Cette hypothèse est confortée par le fait que la même locutrice emploie correctement le pronom *qui* dans d'autres contextes : (KEI/IM/F/dia.) : ah ++ mauvais chose ++ ah ++ il y a + beaucoup des gens + **qui** + sont dans la rue + avec ses chiens + c'est un peu + mauvais (rire)

[5] Dans le corpus, l'emploi du relatif simple *dont* ne se trouve nulle part. Pour ce qui est des relatifs composés, nous avons relevé une seule occurrence, dont l'emploi est biaisé, de 'au(x)quel(les)' ([okɛl]).

[6] Pour le dialogue, le nombre total de mots désigne ceux des locuteurs et non de l'enquêtrice.

Récit de fiction :

	Constructions relatives avec les éléments introducteurs				Relatives simples	Nombre total de mots selon les stades[7]
	il y a	c'est	j'ai	Pseudo-clivé	Autres contextes	
PI	1				3	2033
IM	2				5	2459
AB	1	1			3	2007
AM	8	2			21	3495
Total	12	3			32	9994

Ce tableau montre non seulement l'émergence des structures pseudo-clivées (*ce que/qui…c'est*) à partir du stade IM mais aussi la progression des constructions relatives simples, dépourvues de présentatifs. De plus, l'emploi des constructions à présentatifs est le plus marqué parmi toutes les constructions relatives observées, et notamment dans le stade PI[8]. Les relatives employées dans le stade PI se manifestent donc plus avec ces présentatifs. Nous pouvons schématiser ces constats de la manière suivante :

Relatives-présentatives (*c'est/il y a* X *qui/que* Y)			Relative simple (X *qui/que* Y)
PI	–	>	–
	↓		↓
AM	+	<	+

Aux stades avancés, dans lesquels tous les types de constructions relatives s'observent plus qu'au stade débutant, les structures relatives-présentatives sont moins utilisées que les constructions relatives simples. Ce constat va à l'encontre du stade PI : en effet, les relatives ayant les présentatifs sont moins utilisées par rapport aux autres stades, mais elles sont plus employées parmi toutes les relatives observées. Pour ce phénomène, nous pouvons émettre l'hypothèse suivante : au stade débutant, les pronoms relatifs notamment *qui*, ne sont pas considérés comme un élément grammaticalement indépendant, mais comme un élément s'intégrant dans un patron syntaxique tel que « c'est/il y a…qui/que ». Il en va de même pour la combinaison « j'ai…qui/que », malgré le faible nombre d'occurrences. Tandis qu'au stade avancé, les apprenants peuvent produire librement des relatives sans discrimination de contextes.

Pour ce qui est des relatives simples qui se développent progressivement (tableau 2), l'emploi des pronoms relatifs varie à partir du stade IM dans lequel les apprenants commencent à utiliser *où*. Mais la diversification s'observe principalement chez les apprenants avancés qui tentent d'utiliser d'autres types de pronoms relatifs :

(6) (KAN/AB/F/dia.) : donc **je savais pas trop + la réponse *au(x)quel(les)* je voulais avoir** + et + donc pour ça que j'avais peur pour ++ pour les choses que je connaissais pas

(7) (SAK/AM/F/mono.) : **donc on a reconnu que + c'était elle qui était venue euh au bal et *avec qui* euh le Prince était + tombé amoureux** ++

Malgré l'emploi inapproprié de pronoms relatifs, la flexibilité de leur emploi chez les apprenants avancés semble aller de pair avec la diversification d'autres types d'énoncés complexes.

- **Les constructions syntaxiques observées**

Nous venons d'observer l'emploi des relatives qui sont introduites relativement tôt chez les apprenants japonais. Toutefois, les énoncés complexes les plus précoces dans notre corpus sont caractérisés par la présence de *parce que* et *quand*. Ce fait correspond à ce qui a été observé dans les études antérieures. De plus, comme cela a été signalé par certains travaux sur l'acquisition du FLE (entre autres Bartning 1997 ; Kihlstedt 1998 ; Hancock 2000), notre corpus révèle également un recours progressif aux différents types de énoncés complexes. Nous avons relevé dans le tableau suivant le nombre d'occurrences de différentes constructions et celui de leur variété (indiqué entre parenthèses à droite) : la construction comportant *que*, qu'il s'agisse de la complétive ou de la circonstancielle (*parce que, il m'a dit que, j'espère que…*) et la construction conte-

[7] Pour le dialogue, le nombre total de mots désigne ceux des locuteurs et non de l'enquêtrice.

[8] Ce n'est toutefois pas le cas du récit de fiction. Etant donné le peu d'occurrences dans les stades PI, IM et AB pour le monologue, les chiffres ne semblent pas généralisables et il nous paraît plus pertinent de nous appuyer dans ce cas sur les résultats obtenus dans le dialogue.

nant une proposition interrogative indirecte (*je ne sais pas/je me demande où, comment, si…*).

Tableau 3 : Nombre d'occurrences de la structure complexe et de sa variété[9]

Récit personnel :

	parce que	*quand*	*-que- (complétive)*	Interrogative indirect *(où/comment…*	total
PI	77	30	37 (6)	5 (4)	149
IM	67	39	67 (17)	10 (4)	183
AB	36	25	64 (17)	10 (4)	135
AM	38	24	107 (22)	8 (5)	177
total	218	118	275	33	644

Récit de fiction :

	parce que	*quand*	*-que- (complétive)*	interrogative indirect *(où/comment…*	total
PI	9	3	11 (7)	0	23
IM	6	2	7 (5)	1 (1)	16
AB	3	2	14 (8)	1 (1)	20
AM	7	6	31 (13)	1 (1)	45
total	25	13	63	3	104

L'emploi de l'énoncé complexe est plus fréquent aux stades avancés, notamment dans le stade AM, et cela est plus visible pour le récit de fiction. Comme le montre le chiffre entre parenthèses à droite, la construction complexe varie également de plus en plus : elle est limitée dans le stade PI à la construction comportant « parce que/ quand/ verbes d'opinion+que/ dire que/ vouloir que », tandis qu'elle est plus diversifiée aux stades avancés : l'expression « se rendre compte que… », par exemple, qui n'est pas attestée dans le stade PI ni dans le stade IM commence à être employée dans le stade AB (2 occurrences) malgré le problème de la morphologie verbale (*j'ai [ʁãd] compte que…*) et devient plus fréquente dans le dernier stade (5 occurrences). Par contre, cette diversification de la construction est moins visible pour les propositions interrogatives indirectes : nous trouvons dès le stade PI l'emploi de l'énoncé complexe du type « je ne sais pas *comment faire* », bien que cette construction présente parfois un problème d'ordre syntaxique comme « je ne sais pas *c'est* pourquoi (YAM/PI/F/dia.) ».

La difficulté pour construire l'interrogative indirecte s'observe même dans le dernier stade. Dans l'énoncé ci-après (8), le problème apparaît d'une part dans le manque d'une séquence (*j'ai appelé mon propriétaire* [*pour demander*] *s'il avait un problème…*) ou dans la sélection inappropriée du verbe au contexte (*appeler* au lieu de *demander*) et d'autre part dans la séquence au discours direct/indirect. L'énoncé (9) montre quant à lui le changement fonctionnel du marqueur *si* : il est employé, semble-t-il, au début en tant que marqueur d'interrogation indirecte précédé de « je ne sais pas », mais la dernière proposition (*pourquoi pas déménager au Japon…*) suggère qu'il s'agit d'un « si-hypothétique »[10] :

(8) (TSU/AM/F/dia.) : mais il y avait toujours pas d'eau ++ et du coup **j'ai appelé mon propriétaire + *si* + il avait un problème avec + de la canalisation** […] non parce que j'ai envoyé un mail au proprio **pour lui demander si est-ce que c'est normal** + de ++ ne pas avoir d'eau chau- de ++ l'eau chaude +

(9) (SHO/AM/F/dia.) : **mais je ne sais pas *s'*il y a + un offre d'emploi qui est plus intéressant**

[9] Le comptage s'est effectué sur toutes les utilisations de marqueurs – traditionnellement appelés « conjonctions de subordination ». De ce fait, le marqueur *que* suivi de pause, par exemple, est inclus dans le tableau. Par contre, nous n'avons pas pris en compte l'énoncé incomplet.

[10] Le tableau 3 contient également cette construction. Le nombre d'occurrences de « si-hypothétique » est le suivant : le récit personnel contient 51 occurrences au total (20 occurrences au stade PI, 7 au stade IM, 13 au stade AB et 11 au stade AM). Cette construction est en nombre restreint dans le récit de fiction (1 occurrence dans chaque stade PI, IM et AB et 3 occurrences dans le dernier stade).

pourquoi pas déménager au Japon ça me fait pas peur +

Nous constatons des difficultés sur l'emploi de subordonnées, introduites par *que*, et cela semble caractériser les stades au-delà du stade PI. Les exemples suivants présentent l'insertion inadéquate du marqueur *que* dans des contextes qui ne le requièrent pas. Dans l'énoncé (10), la locutrice introduit *que*, précédé du pronom sujet *je*, initialement prononcé après le mot contenant l'adjectif interrogatif. Cela montre que, malgré la présence de pauses, la locutrice a délibérément ajouté *que*. Quant à l'exemple (11), *que* est placé directement après l'adverbe interrogatif :

(10) (HAM/IM/J/dia.) : hmm ++ pourquoi ++ comment [di] je je **je ne sais pas + hm + quel mot je + *que* + je dois utiliser** mais + eh + co- (rire) + comment [di] ++ hm

(11) (KAN/AB/F/dia.) : je voudr/E/ évoluer ma langue française + **c'est pourquoi *que* + je suis venue + à *T* pour [apʁɑ̃] le français** +

Le marqueur *que* est parfois employé pour une séquence qui pourrait être exprimée à l'infinitif bien que ce phénomène ne soit pas restreint aux apprenants japonais (Blanche-Benveniste 1990 : 54) :

(12) (MIW/IM/F/dia.) : après cinq et dix ans + (rire) + c'est sûr que je parle très bien français (rire) + et **j'espère que je ++ hm + je trouve + je trouve le ++ très bien épouse (rire)** + oui […] il n'y a pas de image mais seulement pour ++ **j'aimerais bien *que* ++ rest/E/ en France** eh comme ça

(13) (MAN/IM/F/dia.) : **mais + j'espère que je veux utiliser ++ hm + le français ++** et ++ je voudrais travailler dans une domaine de + la mode ou + possible publicité ou relation publique

(14) (TOM/IM/F/dia.) : quand je passe avec mes ++ camarades étrangères + je parle en [fʁɑ̃s] on parle en [fʁɑ̃s] français + **je je je me sens + que + je suis heureux**[11] **+**

Comme nous pouvons le constater, ces emplois de *que* qui introduisent une complétive s'observent davantage dans le stade IM à partir duquel la construction complexe se diversifie (tableau 3). De plus, la forme temporelle précédée de *que* est dans la majorité des cas exprimée sous la forme du PRE. Dans les stades avancés, les apprenants recourent à l'infinitif :

(15) (TSU/AM/F/dia.) : et comme je pars + enfin je lui donne l'appart-tement + dans deux jours + et + **j'ai un peu peur de + ne pas pouvoir + récupérer ma caution** (rire) +

(16) (TOG/AM/F/dia.) : **je suis vraiment contente ++ de ++ de travailler enfin de + de pouvoir travailler à la fac pouvoir travailler enfin de ++ pouvoir donner les cours de japonais + euh aux étudiants français + et d'avoir euh ++ les collègues + enfin très sympathiques** +

Toutefois, pour ces stades avancés, nous observons des emplois surprenants de *que* : afin d'ajouter l'information nécessaire, la locutrice KAM ci-dessous emploie la relative introduite par le pronom *que* au lieu de l'exprimer avec un adjectif (par exemple, « des endroits *inconnus* ») :

(17) (KAM/AM/F/dia.) : c'est pas forcément en France + **je suis toujours ++ oui + attirée par ++ quelque part que je connais pas du tout** +

Chez la locutrice suivante, *que* semble être en réalité une forme raccourcie de *parce que* ou *vu que*, étant donné la relation informationnelle des propositions :

(18) (HAT/AM/F/dia.) : c'était pas la frayeur + c'était une plutôt l'inquiétude + mais c'est **il y avait aussi euh ++ la frayeur ++ surtout que + euh je suis née à *T1***

Il est intéressant d'observer qu'au stade débutant, le marqueur de jonction *que* est soit absent (19 et 20) soit associé au verbe antéposé en formant une séquence figée « je pense que » (21). Nous soulignons que la séquence verbale contenant les verbes d'opinion comme « je pense/crois/trouve » se situe le plus souvent en position finale d'énoncés chez les apprenants débutants. Ce qui étaye l'hypothèse du figement pour la séquence « je pense que » dans cet exemple. Par ailleurs, à ce stade où l'acquisition de la morphologie temporelle est en cours de développement et où la forme de base prime sur d'autres formes temporelles, les apprenants expriment le passé en se servant de moyens lexicaux comme nous pouvons le constater dans l'exemple (20) : le SP « au Japon » et le localisa-

[11] Dans l'énoncé (14), la proposition introduite par *que* a une fonction adjectivale.

teur spatial « ici » servent à créer un contraste temporel entre le passé et le présent.

(19) (MIF/PI/F/dia.) : ah > + **je pense + tous les Français + n'est + n'est pas sympa + et + et ++ le gens de Paris + n'est pas sympa (rire)** + oui + mais + le gens d'ici à *T* + est très sympa tout le monde gentil oui +

(20) (YSH/PI/F/dia.): eetto [jap. (euh)] + **au Japon + eh ++ je + je pense je pense en France + il y a beaucoup de vins** + de [jap. (et)] ah + ici + j'ai + je ++ j'ai déjà beaucoup de + j'ai j'ai déjà [bwa] de beaucoup de vins + donc c'est très bon + et c'est très pas cher +

(21) (MAH/PI/F/dia.) : ah ++ quand ++ j'ai + entendu + premier fois le français + **je pense que très très ++ ah ++ beau + joli +**

Un élément grammatical qui tend à être absent au début de l'acquisition est employé de manière superflue aux stades avancés. Ce phénomène semble montrer le statut important de cet élément : l'attention des apprenants sur la présence de l'élément grammatical augmente de plus en plus avec l'avancement dans les stades à tel point qu'il reste ancrer, dans l'esprit des apprenants, comme étant un marqueur se manifestant dans les structures complexes.

3 Conclusion

L'itinéraire du développement syntaxique observé dans cette étude correspond en partie à ce qui a déjà été signalé par les travaux antérieurs sur l'acquisition. Au stade PI, la construction la plus fréquente est la structure simple, caractérisée par la juxtaposition ou la coordination. L'emploi des pronoms relatifs s'observe dès ce stade, mais leur fonction ne semble pas encore assimilée : ils apparaissent davantage avec les éléments introducteurs et les erreurs ou l'hésitation entre *qui* et *que* se manifestent lorsque ces pronoms sont employés de manière autonome. A partir de la comparaison d'emploi entre la structure de focalisation dotées des éléments introducteurs (*c'est/il y a…qui/que*) et la construction relative simple (*-qui/que…*), nous avons constaté que les apprenants débutants tendent à recourir proportionnellement plus à la structure de focalisation qu'à la relative simple. Ce qui n'est pas le cas des apprenants avancés qui emploient davantage la relative simple. Pour ce phénomène, nous pouvons émettre l'hypothèse selon laquelle la prééminence d'emploi de la structure de focalisation au stade débutant découle de la haute fréquence d'emploi autonome des éléments introducteurs *c'est/il y a* et que cette tendance s'estompe dans les stades avancés où les apprenants ont assimilé l'emploi des propositions relatives indépendamment des éléments mentionnés. En ce qui concerne la subordination comme *quand* et *parce que*, dont l'emploi est généralement précoce d'après l'observation des travaux antérieurs sur l'acquisition, elle apparaît également dès le stade PI. Malgré l'émergence des complétives du type *je pense que…*, leur emploi est encore instable compte tenu de l'absence fréquente de l'élément grammatical ou de verbe s'intégrant dans la subordonnée. De plus, à ce stade, l'association du sujet et du verbe qui évoque ce type de construction complétive est placée souvent à la fin de l'énoncé (*…je pense/ j'espère.*). Ce qui nous conduit à supposer que la construction complétive se développe dans un premier temps dans la combinaison de sujet-verbe, comme une séquence figée, à laquelle s'ajoute dans un second temps le marqueur de conjonction. Au stade IM, la coordination est également fréquente mais l'emploi de l'énoncé complexe augmente considérablement. L'utilisation des pronoms relatifs *qui/que* devient pertinente. Toutefois, d'autres types de pronoms relatifs composés ou non ne sont pas encore maîtrisés et les difficultés s'observent également lorsqu'ils produisent un discours indirect. De plus, à partir de ce stade, l'emploi de complétives introduites par *que* devient plus fréquent à la différence du stade débutant dans lequel l'utilisation de *parce que* et *quand* prime davantage. Toutefois, les apprenants intermédiaires tendent à employer la complétive *que* dans les contextes où la séquence peut être exprimée avec l'infinitif. Quant aux stades avancés dans lesquels nous trouvons plus de structures complexes variées, la difficulté portant sur le discours indirect persiste encore. Mais le problème concernant l'emploi de la complétive *que* et de l'infinitif, qu'on a pu observer dans le stade précédent, diminue considérablement. Néanmoins, les marqueurs grammaticaux *qui/que*, qui tendent à être absents au stade débutant, sont employés de manière superflue dans les stades avancés. Ce suremploi des marqueurs semble montrer l'importance accordée par les apprenants, passés par les stades dans lesquels ils ont produit l'emploi idiosyncrasique de ces marqueurs.

Références/References

Blanche-Benveniste, C. (1990). « Un modèle d'analyse syntaxique 'en grilles' pour les productions orales », *Anuario de Psicología, n^{o} 47*, pp.11-28, Facultat de Psicologia Uni versitat de Barcelona.

Bartning, I. (1997). « L'apprenant dit avancé et son acquisition d'une langue étrangère, Tour d'horizon et esquisse d'une caractérisation de la variété avancé », *Aile (Acquisition et Interaction en Langue Etrangère)* 9, pp. 9-50.

Bartning, I. & Schlyter, S. (2004). « Itinéraires acqui sitionnels et stades de développement en français L2 », *French Language Studies, 14*. pp. 281-299.

Granget, C. (2014). « Pourquoi l'acquisition des pro noms est plus simple que celle des articles : apport du japonais L1 dans l'expression de la référence aux entités en françaisL2 », *Con grès Mondial de Linguistique Française*, SHS Web of Conferences 8.

Hancock, V. (2000). *Quelques connecteurs et modali sateurs dans le français parlé d'apprenants avancés, Etude comparative entre suédo phones et locuteurs natifs*, Thèse de doctorat, Université de Stockholm.

Housen, A. & Kuiken, F. (2009). « Complexity, Ac curacy and Fluency in Second Language Acquisition », Applied Linguistics, Decem ber.

Kihlstedt, M. (1998). « La référence au passé dans le dialogue, Etude de l'acquisition de la tempo ralité chez des apprenants dits avancés de français », *Cahiers de la recherche, Dépar tement de français et d'italien*, Université de Stockholm.

Klein, W. & Perdue, C. (1997). « The Basic Vari ety (or: Couldn't natural languages be much simpler?) », *Second Language research* 13 : 4, pp. 301-347.

Perdue, C. (ed.) (1984). *Second language acquisition by adult immigrants: A field manual.* Cross-linguistic series on second language research. Newbury House.

Trévisiol, P. (2003). *Problèmes de référence dans la construction du discours par des apprenants japonais du français, langue 3*, Thèse de doctorat, Université de Paris VIII.

Trévisiol-Okamura, P. (2015). « L'acquisition et l'enseignement des relatives en FLE: regards croisés » in Trévisiol-Okamura, P. & Kahe raoui, M., *Les subordonnées, corpus, acqui sition et didactique*. pp. 103-120. Presses Universitaires de Rennes.

Véronique, D. (dir.) (2009). *L'acquisition de la grammaire du français, langue étrangère*. Paris : Didier.

La langue maternelle et les langues non maternelles connues comme recours pour la communication en Portugais Langue Non Maternelle. Une étude de cas.

Carolina Nogueira-François
Université Lille 3
maria-carolina.nogueirafrancois@univ-lille3.fr

Résumé

Dans cette étude de cas, nous mesurons auprès de deux apprenants l'influence de leur langue maternelle, le français, et de leurs langues non maternelles, dans l'élaboration d'hypothèses sur le portugais dans la communication écrite. Nous analysons l'influence de ces langues sous la forme de stratégies compensant les lacunes de leur apprentissage du portugais. Nous examinons si le statut des langues (LM ou LNM) joue un rôle fondamental qui empêche ou déclenche des stratégies afin de communiquer en portugais. Les résultats nous montrent que les deux apprenants utilisent des langues connues pour communiquer en portugais.

1 Introduction

Le processus d'acquisition de la Langue Maternelle (LM) se fait de manière naturelle et inconsciente, car il suffit d'interagir dans la langue pour l'acquérir. En revanche, l'apprentissage d'une Langue Non Maternelle (LNM) dans un contexte scolaire, nécessite une étude de la langue pour communiquer. Dans cette étude, nous réalisons une étude longitudinale afin d'analyser l'influence de la LM et des LNMs dans la communication de deux apprenants français de Portugais Langue Non Maternelle (PLNM). Par ailleurs, il est utile de rappeler que cette étude est la suite d'une précédente dans laquelle nous avons analysé l'influence de l'espagnol dans 380 productions écrites d'apprenants français de PLNM de l'Université Lille 3 (voir NOGUEIRA, 2014). Les résultats de notre analyse précédente nous ont montré que les étudiants ayant une connaissance préalable en espagnol étaient plus proches de l'apprentissage du PLNM que ceux n'ayant jamais eu de contact avec cette langue. Dans l'étude présente, un seul apprenant parle l'espagnol.

2 Cadre théorique

Opposées au behaviorisme et au structuralisme, les théories cognitivistes de Chomsky (1965, p. 55) sur la prédisposition innée des enfants à acquérir le langage et la séparation entre *Compétence* (la connaissance

que l'on a de la langue) et *Performance* (la communication dans la langue) créent un nouveau paradigme dans le champ d'enseignement apprentissage de LNMs. De même, le concept d'un système abstrait que l'on crée mentalement lorsque l'on initie le processus d'appropriation d'une LNM, baptisé *interlangue* par Selinker (1972), est encore aujourd'hui sujet de nombreuses études. Cependant, s'y ajoutent de nouveaux aspects : psychologiques, contextuels, interactionnistes, entre autres (Gass & Selinker, 2008).

Nous sommes en accord avec Py (2000) lorsqu'il affirme que l'apprentissage d'une LNM et la communication exolingue constituent un effort vers « la construction d'une sorte de rationalité et d'intelligibilité linguistique. Cet effort est davantage un processus qu'un résultat, et l'interlangue se manifeste comme émergence d'une organisation fragile, faite d'une multiplicité hétérogène de microsystèmes instables » (p. 401). Ainsi, dans l'effort de communiquer, l'apprenant peut se servir de stratégies, telles que mélanger les langues ou créer des mots, pour se faire comprendre dans la langue qu'il est encore en train d'apprendre. Dörnyei (1995, p. 56) les nomme *stratégies de communication* dans la LNM.

Si l'utilisation d'autres langues semble être une étape naturelle de l'apprentissage d'une LNM, nous nous demandons quels facteurs favorisent le recours à une langue au détriment d'une autre. Corder (1981, p. 96) explique l'influence de la LM par une extension des habitudes créées dans cette langue et reproduites dans la LNM. En revanche, de nombreuses études soutiennent l'idée que les LNMs connues exercent une influence plus proéminente sur l'interlangue (cf. De Angelis & Selinker, dans Cenoz *et al*., 2001 et Hammarberg, *idem*, 2001). Pour De Angelis & Selinker (*idem*, p. 56), la différence centrale entre l'activation de la LNM au détriment de la LM comme source d'influence est directement liée à leurs statuts : tandis que l'influence d'une LNM provoque la sensation du parler étranger – puisqu'il s'agit de l'interférence d'une interlangue sur l'autre –, l'emploi de la LM n'engendre pas le même effet. Par ailleurs, Hammarberg (*idem*, pp. 22-23) affirme que la proximité typologique entre les langues constitue l'un des principaux facteurs qui engendre l'influence d'un système sur l'autre.

Dans ce contexte, et précisément dans cette étude de cas, nous avons d'abord attiré l'attention des deux apprenants de PLNM sur la proximité des langues romanes. Cette sensibilisation contribue à retirer l'aspect étranger du portugais en soulignant les traits communs entre le portugais, une langue encore inconnue pour eux, et les langues latines qu'ils connaissent : le français (leur LM) et l'espagnol pour un seul sujet (une de ses LNMs). Notre démarche sert non seulement à les encourager à activer les connaissances linguistiques dont ils bénéficient pour communiquer

en PLNM, mais aussi à stimuler les apprenants afin qu'ils développent une motivation pour apprendre le PLNM[1]. Selon Corder (*op. cit*, p. 6), la motivation est l'élément qui peut remplacer la prédisposition des enfants à acquérir le langage.

Nous faisons dans ce qui suit une présentation de notre méthodologie de travail : les sujets, l'objet de notre étude et la méthodologie utilisée pour la récolte et l'analyse des données.

2.1 Méthodologie et analyse

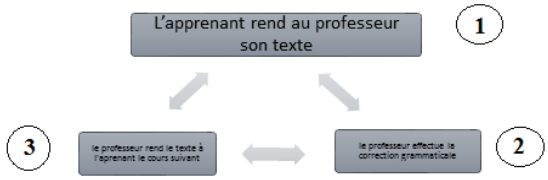
Au sein d'un groupe de sept apprenants d'une grande école française, nous choisissons deux sujets grâce à un questionnaire sur leur connaissance préalable en langues. Voici les informations récoltées du sujet 1 (S1) et du sujet 2 (S2) :

	Âge	LM	LNM1	LNM2
S1	21	français	anglais : 14 années	espagnol : 7 années
S2	21	français	anglais : 12 années	allemand : 5 années

Tableau 1 : Connaissances préalables de LNMs par S1 et par S2

S1 et S2 ont le français comme LM. Ils bénéficient de plus de dix ans d'étude d'anglais. S1 a sept années d'études d'espagnol. S2 a cinq années d'études d'allemand. La connaissance en espagnol a déterminé le choix des sujets : S1, 7 années ; S2 : 0. Les deux apprenants n'avaient aucune connaissance en portugais avant de commencer les cours de PLNM.

Les cours ont eu lieu au rythme de deux heures hebdomadaires, pendant 24 semaines, durant les années scolaires 2014-2015 et 2015-2016. Nous avons constitué notre corpus à partir des productions écrites des sujets. Celles-ci font partie d'une activité écrite non obligatoire, de genre et de thème libres. L'objectif de l'activité est de communiquer en portugais. Au bout de deux années de cours de PLNM, cette activité a engendré 21 textes rendus par S1 d'une part, et 18 textes rendus par S2 d'autre part. Notre corpus se compose de mots et de phrases basées sur le français, l'espagnol et l'anglais.



En ce qui concerne les textes, le mécanisme de récolte des données était le suivant :

Figure 1 : Mécanisme de récolte des textes

1) Les sujets rendent leurs textes au professeur,
2) Le professeur corrige la grammaire,
3) Le professeur rend les textes aux apprenants lors du cours suivant.

[1] Degache (*Des outils numériques pour l'Intercompréhension réceptive*, vidéo, S.D.) soutient qu'une des fonctions de levier didactique de l'intercompréhension est de déclencher la motivation chez les apprenants.

Par ailleurs, nous n'avons utilisé aucun outil sophistiqué pour l'analyse des données, Nous avons procédé à des analyses statistiques simples, basées sur les pourcentages d'occurrence des items relevés. En outre, l'étude longitudinale nous permet de mesurer l'évolution de l'utilisation de la LM et de la LNM comme recours pour communiquer en PLNM. De ce fait, pour faciliter l'analyse, nous avons divisé les textes de chaque sujet en phases selon cette évolution du corpus (augmentation ou diminution du recours à d'autres langues et du nombre des mots par texte). Le résultat de la division est le suivant :

- quatre phases pour S1,
- deux phases pour S2.

Nous identifions le français et l'espagnol comme recourt sous la forme :

- d'emprunts lexicaux (l'utilisation d'une langue dans la communication d'une autre),
- de néologismes (la création d'un nouveau mot : noms, verbes et adjectifs), et
- de calques (transposition d'éléments morphologiques, syntaxiques et morphosyntaxiques d'une langue à l'autre ou la traduction littérale d'une langue dans l'autre).

Passons à l'analyse des phases de nos sujets.

Analyse

Comme nous l'avons indiqué, nous divisons la production des apprenants en phases. Voyons les résultats de notre analyse des textes de S1 et S2 au cours de ces phases.

S1

L'apprenant rend 21 textes au professeur, dans lesquels nous récoltons les données suivantes :

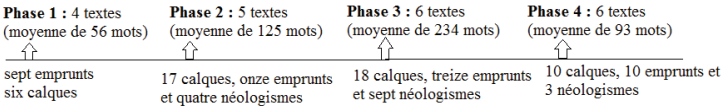


Figure 2 : Les données récoltées des 4 phases de S1

Phase 1 : S1 emprunte à sept reprises (à six reprises du vocabulaire de l'espagnol), comme dans l'exemple suivant : (1) *Chartres é uma ciudad muita bela* (*ciudad # cidade*). En ce qui concerne les calques, S1 débute ses deux premiers textes ainsi : (2) *vou a falar*. Nous considérons la structure (2) comme un calque syntaxique du futur périphrastique espagnol (*voy a hablar*). Notons qu'en français, comme en portugais, le futur périphrastique se réalise de manière similaire (auxiliaire + verbe), sans l'ajout de la préposition *a*.

Phase 2 : S1 élabore plus de calques qu'il n'emprunte de vocabulaire d'autres langues. À titre d'exemple, il transfère vers le portugais le genre masculin de la terminaison française *–age* et de la terminaison espagnole *–aje* : (3) *O piratagem de Game of thrones*. Toutefois, en portugais, la terminaison *–agem* est de genre féminin *(o piratagem # a piratagem)*. Les onze emprunts proviennent de l'espagnol : (4) *E a misma coisa que dos milhãos* (*misma # mesma*).

La phase 3 : cette phase comptabilise le plus grand nombre de mots et de recours à d'autres langues par texte. Les calques y sont les plus nombreux. L'utilisation du pronom

relatif est un exemple de calque syntaxique du français : (5) *A notária quem casou as mulheres* (*la notaire qui*), S1 emploie le pronom relatif portugais *quem* comme on le fait avec le *qui* français (*quem # que*). Les emprunts proviennent tous de l'espagnol : (6) *E asi que se presentaram*. En revanche, dans certains cas, la langue source d'influence des néologismes peut être l'espagnol ainsi que le français, comme par exemple : (7) *se presentaram em frente do juiz para se unir*. Cette forme verbale peut être engendrée par *presentaron* de l'espagnol, ainsi que *présentèrent* du français. Par ailleurs, ce néologisme nous montre le chemin d'une hypothèse validée par le professeur :

phase 3 texte 10	phase 3 texte 11	phase 4 texte 17	phase 4 texte 18	phase 4 texte 18
se presentaram	*apresentaram*	*apresentei*	*apresentar*	*apresentar*

Tableau 2 : Chemin d'une hypothèse sur le portugais basée sur l'espagnol et/ou le français – S1

En parcourant les hypothèses de la construction du verbe portugais *apresentar*, nous constatons que dans le texte 10, l'hypothèse élaborée par S1 (*presentaram*) n'est pas validée par le professeur. Dans le texte 11, S1 suit la correction du professeur et emploie la forme corrigée (*apresentaram*). Dans le texte 17, l'apprenant emploie une autre forme du verbe (*apresentei*). Dans le texte 18, S1 emploie la forme infinitive (*apresentar*) à deux reprises.

Phase 4 : dans cette dernière phase de S1, nous remarquons une réduction du nombre des mots par texte. Le nombre d'emprunts et de calques est identique. Les néologismes continuent à être la stratégie la moins utilisée par l'apprenant. Dans cette phase, les stratégies linguistiques nous montrent la non linéarité de l'élaboration d'hypothèses de l'apprenant : S1 réélabore deux hypothèses non acceptées par le professeur lors de sa première phase (l'emprunt *ciudad* et le futur périphrastique calqué de la syntaxe espagnole *vão a ser secas*).

Cette réutilisation de structures non validées par le professeur nous montre la complexité du processus d'apprentissage d'une LNM. La mémoire joue un rôle prépondérant également dans le processus d'apprentissage : car une hypothèse validée par le professeur ne signifie pas forcément qu'elle sera mémorisée définitivement par l'apprenant.

Nous passons désormais aux résultats de l'analyse de notre deuxième sujet. Se servira-t-il des mêmes stratégies que S1 ?

S2

Comme nous le savons, le fait de n'avoir jamais étudié l'espagnol distingue S2 de S1. Au premier regard, nous constatons l'unique influence du français, sa LM. Observons les données que nous avons récoltées dans les deux phases de S2 :

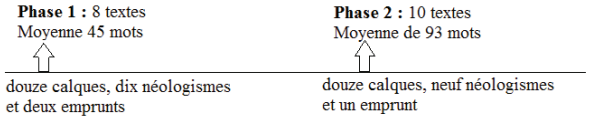


Figure 3 : Les données récoltées des 2 phases de S2

Phase 1 : dans sa première phase, l'apprenant crée plus qu'il n'emprunte. En d'autres termes, il associe des structures du français avec le portugais et crée des formes hybrides inexistantes dans les deux langues : il élabore plus de calques et de néologismes que d'emprunts. À titre d'exemple : (8) *como o revo que realizou.* Le néologisme *revo* (# *sonho*) est la combinaison du mot français *rêve* à la fin duquel l'apprenant ajoute la terminaison de genre masculin en portugais *–o*.

En ce qui concerne ses deux emprunts, à titre d'exemple, S2 utilise un verbe français conjugué à la troisième personne du singulier *dit* pour remplacer le même verbe en portugais *diz*. Notons que les formes se ressemblent (consonne + voyelle + consonne). L'apprenant n'est pas influencé par l'anglais, sa LNM, dans cette première phase.

L'apprenant, dans sa seconde phase, continue-t-il à être influencé uniquement par sa LM pour élaborer ses hypothèses sur le portugais ?

Phase 2 : lors de cette phase, S2 continue à créer des néologismes, comme par exemple

(9) *3 euros e tem 28 étajos*. Ce dernier est créé à partir du nom *étages* pour remplacer le nom en portugais *andares*. Il ne connait probablement pas ce mot en portugais et ajoute la désinence nominale masculine *–o* au nom français *étages*. S2 remplace la consonne /g/ par /j/ pour que le son demeure [ʒ]. Selon nous, il existe déjà une influence du portugais, car S2 aurait pu ajouter la voyelle *–e* pour que le son continue [ʒ], (voir la conjugaison des verbes terminés par *–ger* à la première personne du pluriel, comme par exemple, *nous mangeons*).

Quant à la morphosyntaxe, dans (10) *As melhoras coisas*, la structure peut être un calque syntaxique du français : la traduction littérale de *les meilleures choses*. En effet, cela peut être également dû à une hypercorrection où l'accord est fait : *as melhores coisas*.

C'est seulement dans cette phase de S2 que nous identifions l'influence de l'anglais en tant que stratégie pour communiquer dans notre corpus. La structure (11) *um muro que é 4 metro alto* peut être le résultat d'un calque syntaxique de l'anglais (*4 meters high # 4 metros de altura*).

Nous passons maintenant à nos conclusions, où nous comparons le chemin d'hypothèses élaborées par nos sujets.

3 Conclusion

L'analyse des stratégies de S1 et S2 pour communiquer en portugais nous a permis d'identifier, de mesurer et de comparer l'influence des langues connues dans leur élaboration d'hypothèses sur le portugais. Rappelons que les deux sujets sont francophones et ont plus de dix années d'études d'anglais. S1 a sept années d'étude d'espagnol ; S2 a cinq années d'étude d'allemand. En revanche, nous identifions uniquement le français (S1 et

S2), l'espagnol (S1) et une seule fois l'anglais (S2) dans notre corpus.

Malgré un nombre irrégulier de productions écrites, et de phases[2], la comparaison des huit premiers textes[3] des sujets, nous permet de constater que :

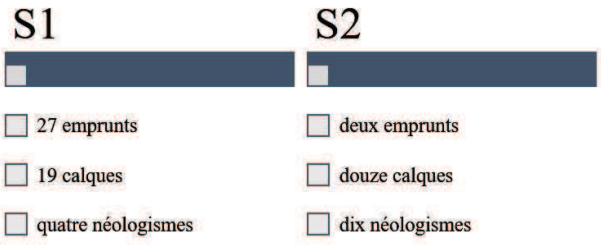


Figure 4 : Les différences entre S1 et S2 dans leurs 8 premiers textes

De plus, suite à l'analyse des productions, nous observons que S2 crée des néologismes à partir de son premier texte. S2 n'élabore le premier néologisme qu'à partir de son cinquième texte. Le français est presque la seule source d'influence de S2 pour compenser les lacunes de son apprentissage du PLNM.

Nous pouvons conclure que les résultats de notre analyse montrent que S1 et S2 ont recours à leurs langues internalisées ou connues comme stratégie pour communiquer en portugais : la LM et la LNM. S1 priorise une LNM (l'espagnol) pour emprunter du vocabulaire : 90% des emprunts proviennent de l'espagnol. Il est possible que l'apprenant ait constaté la ressemblance lexicale entre le portugais et l'espagnol pour avoir privilégié cette langue comme source principale d'emprunts. Cependant, l'apprenant s'inspire de la syntaxe de sa LM. S2, quant à lui, élabore presque toutes ses hypothèses sur le portugais à partir de sa LM, au détriment de l'anglais, sa LNM. D'un côté cela confirme l'hypothèse que l'interlangue est principalement influencée par des systèmes linguistiques typologiquement proches (Hammarberg dans Cenoz *et al*., 2001) – l'anglais étant une langue de la famille germanique, typologiquement plus éloignée du portugais. C'est probablement la raison pour laquelle cette langue n'est pas identifiée dans les textes de S1. De l'autre côté, les résultats de S1 infirment l'hypothèse selon laquelle une autre interlangue peut constituer, dans la communication, une source d'influence plus proéminente que la LM.

Les deux sujets ont donc utilisé la LM et les LNMs pour développer leurs hypothèses sur le portugais : S1, la LNM et la LM ; S2, essentiellement la LM.

Références/References

[ALMEIDA FILHO, J.C.P. (1995) *Português para estrangeiros: interface com o espanhol.*
Campinas : Pontes.
CHOMSKY, N (1965) *Aspects of the theory of syntax*. Massachusetts : MIT Press.

[2] L'irrégularité du nombre de productions des sujets peut être dû au style des apprenants ou même au manque de connaissance d'une LNM plus proche du portugais, comme l'espagnol. La connaissance de l'espagnol semble donner à S1 la sensation de « presque parler » le portugais (Almeida Filho, 1995).

[3] Cette constatation faite dès les huit premiers textes de S1 et S2 se confirme tout au long du corpus.

CENOZ, J. *et al*. (2001) *Cross-linguistic influence in third language acquisition*: *psycholinguistic perspectives*. Great Britain : Cromwell Press Ltd.

CORDER, S. P. (1981) *Error Analysis and Interlanguage.* Oxford : University Press Walton.

DEGACHE, C. (S.D) *Des outils numériques pour l'IC réceptive*. [vidéo en ligne].

DÖRNYEI, Z. (1995) « On the teachability of communication strategies ». Budapest : Université Eötvös, TESOL QUARTERLY, Vol. 29, N°. 1.

GASS & SELINKER (2008) *Second language acquisition : an introductory course*, 3ème ed., New York : Routledge.

NOGUEIRA-FRANÇOIS, C. (2014) L'utilisation des hispanismes en tant que stratégie d'apprentissage du Portugais Langue Étrangère par des étudiants francophones. Mémoire de Master 1. Lille : Université Lille 3.

PY, Bernard (2000) *Didactique des langues étrangères et recherche sur l'acquisition. Les conditions d'un dialogue*. Études de Linguistique Appliquée ; Paris.

SELINKER, L. (1972) « Interlanguage ». *IRAL*, 10:3, pp. 209-230.]

L'alternance modale après les constructions impersonnelles *sembler que* — étude préliminaire statistique à une approche TAL

Divna Petković, Victor Rabiet
Faculté de philologie, Université de Belgrade (Serbie), Université Paris Est (Marne-la-Vallée, France)
didimos88@hotmail.com, victor.rabiet@ens-cachan.fr

Abstract

Dans cet article, nous cherchons à déterminer des paramètres grammaticaux possédant une relation de liaison avec l'alternance modale dans les subordonnées complétives lorsque celle-ci intervient après les constructions impersonnelles avec le verbe *sembler*. Établis sur un petit échantillon (étiqueté manuellement), ces paramètres, une fois caractérisés, ont pour vocations de permettre une exploration à grande échelle de manière automatisé : une perspective TAL, par exemple dans le cadre d'une application à l'amélioration de la traduction automatique du subjonctif, ou encore dans une meilleure compréhension de la tendance de textes analysés automatiquement, peut commencer à être envisagée.

1 Introduction

Dans l'article (Petković and Rabiet, 2016) nous avions abordé la problématique de l'alternance modale à travers le spectre de la polysémie, en utilisant deux approches distinctes, celle de Soutet et de Victorri.

Nous avions alors fourni une liste de verbes répondant à ce schéma et pouvant, au moins pour certains d'entre eux, donner un espoir de désambiguïsation grâce à l'alternance modale.

Il est alors apparu que, pour espérer arriver à une telle fin, il fallait, a priori, mener une étude individuelle d'un certain nombre de ces verbes. Nous avons donc décidé de nous intéresser ici au verbe *sembler* et, plus précisément, à la construction impersonnelle

Il <sembler> que

Notre but dans cet article est de faire une étude préliminaire concernant les facteurs montrant une corrélation [1] à l'utilisation, dans la complétive, du mode subjonctif plutôt qu'un autre (presqu'exclusivement, dans les exemples de notre corpus, l'indicatif). En effet, un outil prometteur dans l'analyse des différents facteurs influant sur ce choix est un outil statistique, classique dans le domaine médical, appelé régression logistique.

Pour mettre en place une telle analyse, nous devons identifier, premièrement, un certain nombre de facteurs d'intérêt potentiels. De plus, pour déterminer l'influence de ceux-ci, il faut disposer de corpus suffisamment grands et déjà annotés selon ces facteurs, et qui, pour permettre un travail efficace et réaliste, doivent être mis en place d'une manière automatisée. C'est ici que notre étude préliminaire prend tout son sens : avant d'établir ce type de corpus de grandes tailles avec des annotations personnalisées selon les besoins des diverses études [2], il est primordial de définir les *potentiels* paramètres d'intérêts. Ce que nous ferons sur un corpus test réduit et que nous illustrerons ici, successivement, essentiellement sur les deux paramètres suivants :

- le temps du verbe *sembler* dans la principale ;
- le temps du verbe dans la complétive

2 Présentation de la problématique linguistique

2.1 Le point de vue de Soutet

Nous retrouvons le verbe sembler dans une construction impersonnelle, qu'on pourrait schématiser en utilisant, à l'instar de Soutet (Soutet, 2000, p. 74-75), la terminologie de Tesnière :

1. Dans tout cet article, *corrélation* n'est pas entendu au sens de *corrélation statistique*, mais au sens courant, à savoir au sens de *relation/liaison de dépendance*. Plus précisément nous dirons ici que les paramètres sont corrélés s'ils n'ont pas une relation d'indépendance entre eux.

2. La réalisation de tels types de corpus est un travail en cours, conjoint avec Philippe Gambette, déjà bien avancé et dont l'exploitation devrait arriver d'ici les prochains mois.

la structure actantielle du verbe divalent *sembler* est *sembler* ***y z***, dont :

y - objet indirect renvoyant à un animé humain

z - forme propositionnelle

Soutet remarque que « [l]e jeu modal dans **z** est ici fortement conditionné par la présence (ou l'absence) de **y**. Aussi bien, si l'absence de **y** concourt fortement à l'emploi du subjonctif dans **z** (Il semble que Pierre parte), sa présence, en revanche, impliquant la prise en charge par une personne de l'« apparence » que signifie le verbe *sembler*, favorise nettement le mode indicatif (Il me semble que Pierre part). »

Pour vérifier plus concrètement ce que l'on entend par « *concourir fortement* » et « *favoriser nettement* », nous avons, dans un premier temps, effectué une analyse préliminaire [3] : parmi les 76 résultats obtenus, les 4 constructions impersonnelles avec le verbe *sembler* étaient présentes dans (Petković and Rabiet, 2016), dont [1] *sembler que* + subj. (13 occurrences), [2] *sembler que* + ind. (4), [3] *sembler que* + COI + ind. (57), [4] *sembler que* + COI + subj. (2).

Suite à cette expérience, nous avons voulu faire une étude statistique plus complexe, cherchant les ratios d'occurrence pour chaque cas de figure.

2.2 Les remarques dans le *Bon usage*

On se reportera à (Grevisse, 1975, p. 1454-5).

Grevisse et Goosse affirment le fait suivant : « *Quand* sembler *pris affirmativement est accompagné d'un objet indirect, on met le plus souvent l'indicatif* [...] *Le subj. se trouve pourtant dans la langue littéraire* ».

D'un autre côté, lorsqu'il s'agit de la construction *il semble que*, si « *ce verbe pris affirmativement n'est pas accompagné d'un objet indirect, on met l'indicatif ou le subjonctif.* » Nous remarquerons, donc, qu'aucun des modes n'est considéré comme plus courant que l'autre (contrairement à ce que dit Soutet).

2.3 Étude statistique de B. Hasselrot

Grevisse et Goosse citent aussi un article très intéressant de Bengt Hasselrot, publié dans la Revue romane, 1973, (Hasselrot, 1973, pp. 70-80), qui constate que le subjonctif est plus fréquent après *il semble que* qu'après *il semblait que*.

Nous retiendrons plusieurs remarques pertinentes de cette étude, notamment celle sur l'importance du registre, qui peut être considérable, ce que Hasselrot montre en comparant son corpus (85 % d'exemples provenant de la presse des années 1970-1971) et celui de H. Nordahl, qui comprend 156 romans du XXe siècle (Nordahl, 1969). Nous reviendrons sur le problème du registre dans la section Les paramètres d'intérêt.

Lorsqu'il étudie les exemples de son corpus ou semble que est suivi de l'imparfait ou du passé simple, il est, selon ses propres mots (Hasselrot, 1973, p. 72), en accord avec Boysen qui constate que l'imparfait et le passé simple expriment « une nuance aspectuelle que le subjonctif ne rend pas. » (Boysen, 1971, p. 30). Ceci reste une piste a explorer dans le futur, dans le cadre de nos recherches ultérieures.

3 Corpus et méthodologie

Nous avons effectué nos recherches dans le corpus Frantext `www.frantext.fr/`, base textuelle de référence. On a choisi la période la plus contemporaine, entre les années 2000 et 2016, pour donner un aperçu de la situation actuelle dans la langue, sachant que l'on s'attend de moins en moins à trouver les subjonctifs, même dans les œuvres littéraires, et partant de l'hypothèse que cette réalité linguistique pourrait éventuellement changer la donne dans les cas de l'alternance modale. Ceci est, donc, une contribution de "mise a jour" aux études qui existent déjà sur cette question.

Le seul problème mineur de cette approche, comme nous verrons plus tard, se trouve dans les dates d'édition de certains ouvrages, qui sont, en fait, des rééditions ou des œuvres complètes, mais cela est un manquement de Frantext qui n'indique pas les dates des premières éditions, et il faudrait y pallier dans un article beaucoup plus détaillé.

3.1 Description du corpus de travail

Notre corpus de travail est basé sur la recherche de tous les textes de la période 2000-2016 dans Frantext.

Recherche dans un élément bibliographique : 2000-2016 (dans la date)
Nombre de textes : 188
Nombre de mots : 14 334 553

---

3. À l'aide du Corpus parallèle français-serbe de 1 000 000 de mots — de textes littéraires depuis 1850, `http://www.korpus.matf.bg.ac.rs/`.

Dans le corpus de travail ainsi créé, nous avons cherché l'expression de séquence suivante [4] :

il &q(0,1) &csembler &q(0,7) que

Cette recherche a fourni 821 résultats.

3.2 Dépouillement de la concordance

Pour obtenir les résultats non-ambigus, nous avons été obligés de trier la concordance en éliminant les cas suivants :

- *sembler* + infinitif

  parce que l'infinitif remplace la complétive. Par ex :

(**1**) *Je suis retournée à l'école le jeudi. Il me semblait devoir retrouver au plus vite ce lieu que j'avais jusque-là mis tant d'énergie à fuir.*
(BOULOUQUE Clémence, Mort d'un silence, 2003, 109-111)

- *sembler* + adjectif

  Par ex :

(**2**) *Mais son père semblait si heureux que son nouveau tracteur soit mis à l'honneur...*

- en incise

  Par ex :

(**3**) *C'est la première fois, il me semble, que j'ai le sentiment, violent, comblant, de jouer.*
(OZOUF Mona, Composition française : retour sur une enfance bretonne, 2009, 104-105)

- *il semble que oui/non*

  Par ex :

(**4**) *L'école de l'Église où s'annoncera, la fin du monde venue, le jugement général ? J'aimerais pouvoir le penser, mais il semble bien que non, car nous devons répéter aussi que « l'Enfer est un lieu dont on ne sort jamais ».*
(OZOUF Mona, Composition française : retour sur une enfance bretonne, 2009, 140-141)

- les homographies

  Par ex :

(**5**) *Allons-y, je soupire,inquiet malgré tout de cette intrusion nocturne. Car il semble que, depuis le lointain début de notre soirée dînatoire, nous digressions par successifs et insidieux paliers vers des zones à hauts risques, pousserons-nous l'imprudence à son comble.*
(GARAT Anne-Marie, Programme sensible, 2012, p. 179)

4 Analyse du corpus

4.1 *il <sembler> que* avec COI

Nous avons trouvé 426 occurrences d'indicatif après la construction *il semble que* + COI + indicatif, et seulement 9 occurrences de subjonctif — toutes les 9 chez Marcel Aymé. Étant donné que l'œuvre de Marcel Aymé date des années 50, nous pouvons constater qu'aucun exemple de subjonctif après *il semble que* + COI n'a été trouvé. Néanmoins, ce résultat n'est pas décourageant, car il nous laisse une ouverture pour une recherche ultérieure, soit dans un corpus beaucoup plus large, couvrant le XX^{e} siècle en entier, soit dans l'œuvre d'Aymé elle-même, ce résultat pouvant éventuellement indiquer une marque de son style littéraire.

4.2 *il <sembler> que* sans COI

Après le dépouillement de la concordance, nous avons trouvé 171 occurrences de la construction *il <sembler> que* sans complément d'objet indirect, dont :

- 74 avec l'indicatif dans la complétive
- 97 avec le subjonctif dans la complétive

5 Les paramètres d'intérêts

Comme nous l'avons évoqué dans l'introduction, notre objectif est d'arriver à caractériser ou identifier les éléments permettant de prédire, avec plus ou moins de précision, l'alternance modale.

La première étape est donc de définir les paramètres potentiels susceptibles d'influencer ce phénomène. Ces paramètres ne sont pas fixés a priori et peuvent être véritablement quelconques, sans informations additionnelles. On peut imaginer, par exemple, les paramètres suivant :
la période temporelle, le genre littéraire, le niveau de langue, des caractéristiques grammaticales dans la principale ou la subordonnée, etc…

L'intérêt d'une étude préliminaire est donc d'identifier ceux qui représentent potentiellement un véritable intérêt.

4. Pour obtenir les résultats sans et avec COI, nous avons pris en compte le fait que le COI peut se trouver après [sembler] (*Il semblait à Paul...*), mais avant *que*, aussi bien que les modificateurs adverbiaux (bien, en effet...).

Par exemple

- la *période temporelle* est un facteur intéressant (même si, a priori, on peut imaginer intuitivement que plus l'on se rapproche de la période classique, plus l'usage du subjonctif est fréquent), mais il n'est ni facile à définir (fractionner le temps en tranche de périodes de quelles longueurs ? régulières ? suivant des événements historiques ? littéraires ?) ni à caractériser : il y a en effet le phénomène de réédition qui, par exemple sur Frantext, influence les données (et les multiplies éventuellement), les éditions post-mortem, etc…

- le *niveau de langue* (qui également, semble, déjà a priori, lié à la fréquence de l'usage du subjonctif) qui est intuitif, demande une caractérisation manuelle ***et*** réalisée par un locuteur natif et n'est donc pas directement utilisable dans une perspective de TAL.

Nous observerons ici essentiellement les deux paramètres suivants, pour donner une idée de notre méthode :

- le temps du verbe *sembler* dans la principale ;

- le temps du verbe dans la complétive

6 Tests statistiques et premiers résultats

6.1 Corrélation [5] temps dans la principale — alternance modale

Nous parlons ici de l'alternance modale en tant qu'alternance entre le mode indicatif et subjonctif : nous omettrons donc les 5 exemples de notre échantillon possédant un mode conditionnel dans la complétive (nous ferons une remarque concernant l'éventuel ajout de ses exemples omis à la fin de cette sous-section).

En omettant également les exemples ayant une principale au conditionnel (17 cas), nous obtenons le tableau de fréquence suivant (où *subj.* et *ind.* indiquent le mode de la complétive et où la colonne de gauche indique le temps (à l'indicatif) de la principale) :

	subj.	ind.
passé	18	24
présent	66	41

FIGURE 1 – Temps principale v.s. mode complétive

sur lequel on effectue un test du χ^2 et qui nous fournit une « p-value » $p = 0,037$ ce qui nous permet de rejeter l'hypothèse d'une indépendance des variables **temps (dans la principale)** et **mode (de la subordonnée)** avec une forte présomption contre celle-ci (seuil inférieur à 5%). Il y a donc une corrélation significative [6] sur cet échantillon.

Remarque 1 *La probabilité de cette « corrélation »* (i.e. *du rejet de l'hypothèse d'indépendance des variables) semble augmenter avec la taille de l'effectif, ce qui est bon signe (il est important de garder en tête que, plus l'échantillon est petit, moins le test est fiable, le test du χ^2 étant un test asymptotique) : initialement nous avions observé le sous-corpus concernant la période 2000-2009, ce qui donnait également une corrélation mais avec une certitude plus faible, de l'ordre de* 80%.

Si l'on intègre les exemples dont le mode dans la principale est le conditionnel, nous obtenons le tableau

	subj.	ind.
passé	18	24
présent	79	45

FIGURE 2 – Temps principale (avec cond.) v.s. mode complétive

ce qui donne un résultat encore meilleur avec $p = 0,017$, soit une « corrélation » (rejet de l'hypothèse d'indépendance) avec un seuil inférieur à 2% (le cas du sous-corpus donnait également une meilleure certitude lorsque l'on prenait en compte les conditionnels dans la complétive que lorsque l'on les omettait).

Cette remarque concernant l'ajout des exemples possédants une principale au conditionnel présent (tous les exemples conditionnels étant au présent), nous pousse à nous interroger brièvement sur l'éventuelle corrélation entre le fait d'avoir une

5. Cf. note 1.

6. De même, dans tout cet article, *corrélation significative* est entendu au sens qu'il y a une forte présomption contre l'hypothèse d'indépendance des deux variables.

principale au présent de l'indicatif ou du conditionnel et une subordonnée au subjonctif ou à l'indicatif. On obtient ici le tableau fréquentiel suivant :

	subj.	ind.
présent ind.	66	35
présent cond.	13	4

FIGURE 3 – présent (ind./cond.) principale v.s. mode complétive

Avec une *p-value* de $0,419$ (cette fois en utilisant le test exact de Fisher, du fait du petit nombre de certaines observations), la corrélation ne peut être reconnue comme probable et, au moins en première approximation, on ne peut rejeter ici l'hypothèse d'indépendance, ce qui justifie l'ajout ayant donné lieu au deuxième tableau.

Conclusion partielle 1 *On observe une dépendance très probable entre*

- *principale au présent et présence majoritaire de subjonctif dans la subordonnée ;*
- *principale au passé et présence majoritaire d'indicatif dans la subordonnée.*

Remarque 2 *Dans le cadre de cette étude préliminaire, sur ce paramètre, nous obtenons une conclusion plus précise que celle de Soutet (pour qui le subjonctif est dominant après* il <sembler> que *sans COI) : si celle-ci n'est, globalement pas contredite, en travaillant sur un sous-groupe (principale au passé), une situation paradoxale semblerait apparaître par rapport à la « règle » globale énoncée par celui-ci.*

Précisons également que notre résultat est en accord avec (Hasselrot, 1973).

6.2 Corrélation temps dans la subordonnée — alternance modale

En omettant — dans la principale — les 17 occurrences de conditionnels et — dans la subordonnée les 5 conditionnels ainsi que les 8 futurs, on obtient le tableau

	subj.	ind.
passé	43	45
présent	41	13

FIGURE 4 – Temps complétive v.s. mode complétive

ce qui nous donne une *p-value* de $0,0014$, et par conséquent une dépendance extrêmement probable entre le temps de la complétive et le mode.

Pour évaluer l'influence potentielle de la concordance des temps, observons les exemples où celle-ci n'a pas d'impact sur le choix du temps de la subordonnée.

Pour cela nous excluons les exemples correspondant à un passé dans la principale :

	subj.	ind.
passé	25	22
présent	41	13

FIGURE 5 – Temps complétive (sans principale au passé) v.s. mode complétive

ce qui donne une *p-value* de $0,016$, et donc un seuil de rejet de l'hypothèse d'indépendance du même ordre que pour le test entre indépendance entre *temps de la principale* et *mode de la complétive*, et qui semble bien montrer l'impact de la concordance des temps.

En ajoutant les 16 occurrences de conditionnels compatibles (on exclut celle associée au futur dans la subordonnée) dans la principale, on obtient le tableau de la figure 6, qui correspond à une *p-value* de $0,0019$, ce qui ne change quasiment rien de la conclusion obtenue avec le tableau de la figure 4.

	subj.	ind.
passé	44	53
présent	43	18

FIGURE 6 – Temps complétive (inclut principale au cond.) v.s. mode complétive

Conclusion partielle 2 *On observe une dépendance extrêmement probable entre*

- *le temps (passé ou présent) de la complétive et son mode ;*
- *cette probabilité de dépendance semble être encore augmentée par le phénomène de* concordance des temps.

6.3 Remarques

Nous avons considéré l'opposition *passé-présent*, en regroupant tous les temps passés (imparfait, plus-que-parfait, etc…) car certains, au

vu de notre échantillon, étaient fortement minoritaires. Notons également que l'imparfait était le temps largement majoritaire.

Cependant, sur un échantillon plus grand, il n'est pas exclu de penser que l'étude plus précise de certains temps pourrait fournir des paramètres intéressants également.

La validité des résultats, indépendamment de la taille relativement modeste de l'échantillon, dépend également de la représentativité du corpus choisi, ce qui dépend, dans notre cas, du choix des textes retenus dans Frantext. Pour tester dans une certaine mesure (bien que cela ne soit évidemment pas suffisant dans l'absolu !) cela peut être intéressant de tester l'hypothèse d'une corrélation avec un paramètre a priori indépendant de l'alternance modale. Nous avons donc essayé sur notre échantillon le test de l'indépendance entre le sexe des auteurs (qui a priori n'impacte pas de manière évidente l'usage ou non du subjonctif !) et l'alternance modale :

	subj.	ind.
F	40	36
M	57	38

FIGURE 7 – Test sur paramètre a priori indépendant

avec une *p-value* de $0,33$, on accepte l'hypothèse d'indépendance (il est probable que sur un échantillon plus grand, la *p-value* serait encore plus grande, et donc l'indépendance encore plus certaine), ce qui est donc un indice de la bonne représentativité de l'échantillon.

7 Perspectives

Comme indiqué dans l'introduction, l'objectif, à terme, est d'obtenir un dispositif prédictif de l'alternance modale, permettant d'aider à la désambiguïsation pour les verbes possédant une polysémie liée à cette alternance (en plus d'une explication éventuelle de certains phénomènes grammaticaux associés), selon la liste établie dans (Petković and Rabiet, 2016). Ce qui pourrait, par exemple, aider à la traduction automatique, ou encore, mieux repérer la tendance (en terme de signification globale) d'un texte dans le cadre du TAL.

Ainsi, dans un premier temps, nous nous intéressons à la détermination de plusieurs paramètres et à l'étude de l'importance de leur impact respectif sur l'alternance modale : la perspective intermédiaire est d'employer pour cela une régression logistique [7] en faisant varier des bases d'exemples de grandes tailles issues de différents corpus de départ [8]. Pour cela il est important de pouvoir utiliser une extraction automatique de la structure grammaticale qui nous intéresse et de pouvoir étiqueter également automatiquement les paramètres d'intérêt.

L'outil informatique « adapté » est en cours de développement et sera testé sur la suite logique de cette étude, dans un premier temps sur l'ensemble de Frantext, et, dans un second temps, sur des corpus divers. Précisons cependant ce que l'on entend par « adapté » : il est conçu pour pouvoir récupérer des fichiers de type texte, les étiqueter syntaxico-grammaticalement grâce au logiciel UNITEX, récupérer les exemples correspondant à notre schéma (cette fois par un outil de « graphe » également présent dans UNITEX) et générer un fichier de sortie de type tableur avec les exemples qui correspondent à notre étude. Les paramètres qui peuvent être ainsi automatiquement caractérisés, sont (en plus des paramètres éventuellement déjà étiquetés, selon les bases de données de départ utilisées, tels dates, styles, etc…) des paramètres syntaxiques et/où grammaticaux.

8 Conclusion

Dans cet article, nous avons cherché à illustrer un processus de recherche de paramètres en corrélation avec l'alternance modale, lorsque celle-ci apparait après la construction *il <sembler> que*. Cette recherche commence sur de petits échantillons, étiquetés manuellement, à l'aide de tests statistiques élémentaires dans l'objectif de se diriger vers la constitution d'échantillons de grandes tailles de façon informatisé (au moins dans le cas de paramètres syntaxiques où grammaticaux) et permettant ensuite une étude statistique plus poussée, comme, par exemple, la régression logistique.

De nos exemples de paramètres exposés ici, il est ressorti que les paramètres

1. temps du verbe *sembler* (regroupé en deux classes : *passé* et *présent*) dans la principale ;

2. temps du verbe dans la complétive (regroupé en deux classes : *passé* et *présent*)

7. Pour avoir une bonne idée de cette méthode statistique, on pourra consulter le livre en ligne (Rakotomalala, 2011).

8. Pour un exemple d'utilisation d'une telle méthode sur une base d'exemples de petite taille (environ 500) relativement à l'alternance modale voir l'excellent article de Olaf Mikkelsen (Mikkelsen, 2016).

sont en corrélation significative (au sens d'une dépendance significativement probable) avec l'alternance modale (très significative pour le second, avec une influence marquée de la concordance des temps) et méritent, par conséquent, a priori, de figurer comme paramètre d'intérêt dans une étude plus large.

References

Pascal Amsili and Floriane Guida. 2014. Vers une analyse factorielle de l'alternance indicatif/subjonctif. In *SHS Web of Conferences*, volume 8, pages 2313–2331. EDP Sciences.

Gerhard Boysen. 1971. Subjonctif et hiérarchie, étude sur l'emploi du subjonctif dans les propositions complétives objets de verbes en français moderne, études romanes de l'université d'odense.

Ferdinand Brunot. 1922. *La pensée et la langue : méthode, principes et plan d'une théorie nouvelle du langage appliquée au français*. Masson et cie.

Jacques Cellard. 1996. *Le subjonctif : Comment l'écrire ? Quand l'employer ?* De Boeck Supérieur.

Marcel Samuel Raphaël Cohen. 1965. *Le subjonctif en français contemporain : tableau documentaire*. Société d'édition d'enseignement supérieur.

Laurent Gosselin. 2010. *Les Modalités en français*. Amsterdam-New York, Rodopi.

Maurice Grevisse. 1975. *Le bon usage : grammaire française, avec des remarques sur la langue française d'aujourd'hui*. J. Duculot.

Gustave Guillaume, Roch Valin, WH Hirtle, and André Joly. 1971. *Esquisse d'une grammaire descriptive de la langue française (III) et Sémantèmes, morphèmes et systèmes : 1944-1945, Séries A et B. 11*. Presses Univ. Septentrion.

Gustave Guillaume. 1992. *Esquisse d'une grammaire descriptive de la langue française (III) et Sémantèmes, morphèmes et systèmes : 1944-1945, Séries A et B. 11*. Presses de l'Université Laval, et Lille, Presses universitaires de Lille.

Bengt Hasselrot. 1973. Répartition des modes après' il semble que'essai de statistique linguistique comparée. *Revue romane*, 1.

Eva Havu. 1996. *De l'emploi du subjonctif passé*, volume 285. Helsinki, Academia Scientiarum Fennica.

Paul Imbs. 1953. *Le subjonctif en français moderne : essai de grammaire descriptive*, volume 11. Faculté des Lettres de l'Université de Strasbourg.

Robert Martin. 1983. *Pour une logique du sens*. Paris, PUF.

Robert Martin. 1990. Pour une approche vériconditionnelle de l'adverbe" bien". *Langue française*, (88) :80–89.

Olaf Mikkelsen. 2016. Libre choix de mode ? Vers une analyse multifactorielle de l'alternance indicatif/subjonctif en français contemporain. *HAL*.

Henning Nølke. 1985. Le subjonctif : fragments d'une théorie énonciative. *Langages*, (80) :55–70.

Henning Nølke. 1994. La dilution linguistique des responsabilités : Essai de description polyphonique des marqueurs évidentiels" il semble que et il paraît que". *Langue française*, pages 84–94.

Helge Nordahl. 1969. *Les systèmes du subjonctif corrélatif*. Universitetsforlaget.

Divna Petković and Victor Rabiet. 2016. La polysémie lexicale et syntaxique de l'alternance modale indicatif/subjonctif–perspectives TAL. *PARIS Inalco du 4 au 8 juillet 2016*, pages 80–94.

Ricco Rakotomalala. 2011. *Pratique de la Régression Logistique*.

Olivier Soutet. 2000. *Le subjonctif en français*. Ophrys.

Knud Togeby. 1966. La hiérarchie des emplois du subjonctif. *Langages*, (3) :67–71.

Bernard Victorri. 1997. La polysémie : un artefact de la linguistique ? In *Revue de sémantique et pragmatique*, number 2, pages 41–62.

Duško Vitas and Cvetana Krstev. 2006. Literature and aligned texts. *Readings in Multilinguality*, pages 148–155.

Duško Vitas, Cvetana Krstev, and Eric Laporte. 2006. Preparation and exploitation of bilingual texts. *Lux Coreana*, 1 :110–132.

Harald Weinrich. 1989. *Grammaire textuelle du français*. Editions Didier.

Marc Wilmet. 2010. *Grammaire critique du français*. Duculot.

Paramètres prosodiques et ratificationnels au sein des séquences contributionnelles et modélisation de l'interface sémantique/pragmatique

Camille Létang
Université d'Orléans, France

Abstract

Cet article a pour objectif de montrer le double intérêt pour la pragmatique et la modélisation de l'interface sémantique/pragmatique d'une approche des contributions passant par l'étude empirique des mécanismes de ratification contributionnelle, et d'une compréhension élargie du rôle très important que jouent les contraintes de ratification. Est abordé en particulier le rôle de la prosodie, et ceci à la fois dans l'interprétation de l'orientation argumentative de ce qui est dit, et dans la structuration et l'explication des échanges, la prosodie s'avérant être à la fois une marque linguistique explicite - contribuant en cela à modeler le contenu qu'il faut bien appeler sémantique de ce qui est dit - et être à l'origine d'une très grande part de l'interprétation pragmatique des séquences contributionnelles et dialogales, qui jusqu'ici était présumée être totalement implicite.

Introduction

L'objet de cet article est de montrer la façon dont la compréhension de l'interface sémantique/pragmatique, mais aussi de l'interface entre sémantique/pragmatique d'une part et prosodie d'autre part, est éclairée par la transformation de l'étude d'un objet initialement purement théorique, la notion de contribution, en étude empirique des mécanismes de ratification des contributions et de ce fait de « séquences contributionnelles ».

Les travaux de recherche que nous avons mené depuis quelques années portent donc sur la notion de séquence contributionnelle, autrement dit sur l'existence d'une séquence conversationnelle réunissant la contribution (au sens de Nemo 1999, 2007) initiale et l'ensemble du feed-back interlocutif auquel celle-ci donne lieu dans le cadre du processus de ratification qu'enclenche automatiquement toute intervention contributionnelle. Il s'avère en effet que l'étude empirique des contributions ne peut être séparée de la réaction qui lui est associée.

Tout ce que nous avons vu conduit à interroger la théorisation du dialogue, en cela notamment que celle-ci doit intégrer la relation qui existe entre la nécessité pour toute contribution de contribuer à une co-définition de ce qui doit être pris en compte, et la nature des enchaînements discursif (principalement en contexte dialogal). Nous avons tenté d'expliquer comment, sémantiquement et pragmatiquement, se construisent les conversations et comment est influencée l'interprétation, en soulignant pourquoi le rôle que joue la prosodie est primordial en analyse conversationnelle.

1 Les contributions

Les dialogues sont constitués d'interventions/contributions et l'étude de la ratification contributionnelle est à la fois un moyen de comprendre la logique contributionnelle et surtout d'éclairer une dimension du dialogue qui n'a sans doute pas été prise en compte suffisamment, notamment dans sa capacité à structurer les conversations, et d'appréhender une certaine logique des enchaînements dialogaux. Nous nous intéressons à l'existence d'un feed-back contributionnel comme trace des contraintes qui régissent les contributions, et à la façon dont l'étude de la dimension ratificationnelle des échanges contribue à éclairer la dynamique et la structuration du dialogue.

Et ce aussi bien :

— en termes d'explication de la nature des échanges eux-mêmes, et donc des enchaînements dialogaux ;
— au travers de la dimension prosodique de la ratification/non-ratification et de son rôle dans les élaborations dialogales ;
— au travers d'une classe assez largement spécifique (et lexicalisée) de « mots de discours » porteurs de commentaires méta-contributionnels et méta-ratificationnels.

En analyse de la conversation, le terme de *contribution* est généralement employé pour désigner la participation d'un locuteur à une conversation. D'une manière plus spécifique, la définition pouvant être faite en contexte dialogal, car c'est dans ce contexte que les recherches sont ciblées, c'est ce que chacun dit à propos de ce qui doit être pris en compte par tous et par exemple sur ce qui a été dit antérieurement (au sein d'une même discussion). Une contribution peut être produite par un même locuteur sur son propre discours. Mais nous pouvons relever le cas de la contribution polyphonique : quand un locuteur X vient ajouter une information (son intervention remplit la maxime de quantité de Grice), ou tente d'attirer l'attention sur un point. La contribution peut servir à revenir sur des propos antérieurs (par manque d'informations ou stratégies argumentatives), ou à souligner à notre interlocuteur qu'il oublie de dire quelque chose, et ainsi compléter l'énoncé (ce qui peut aller contre la face positive de l'interlocuteur). Cette contribution, insérée dans un énoncé d'un autre locuteur, construit l'échange au plus près de ce sur quoi on veut attirer l'attention. On cherche à respecter la contrainte de dire ce qui parait important, au plus près de la réalité, dans le souci de minimiser l'effort de compréhension de l'interlocuteur.

2 Ratification des contributions

En fonction du contexte, de l'enjeu, les interlocuteurs font le choix d'orienter leurs propos soit en fonction de l'implicite (prosodie), soit en fonction du sens du contenu, des mots employés. Tout dépend de la façon dont on attire l'attention de notre interlocuteur, sur ce qu'on veut qu'il prenne en compte ou non (forme d'argumentation masquée).

Toute contribution peut a priori être :

— rejetée (comme hors de propos ou inacceptable) ;
— ignorée (comme hors de propos ou inacceptable) ;
— ratifiée tacitement par un silence ;
— ratifiée comme secondaire ou marginale, autrement dit comme méritant peu d'attention ;
— ratifiée comme importante, et donc comme méritant d'être prise en compte avec toute l'attention nécessaire.

Sur cette base, nous appelons « ratification » la forme de feed-back que reçoit toute contribution et en fonction de laquelle la demande de prise en compte de quelque chose est intégrée ou non au « consensus interlocutif ». Toute étude empirique du processus de (non) ratification, qu'elle soit menée sur des données orales (par exemple un débat) ou écrites (par exemple un compte-rendu), commence par un repérage sur corpus de tous les éléments (lexicaux, discursifs, prosodiques) ou de toutes les séquences linguistiques qui portent sur des contributions antérieures et sont susceptibles de définir la valeur contributionnelle.

Avant toute forme de contribution, entre en jeu la définition du champ attentionnel, et il faut noter qu'il n'est pas possible de supposer qu'il résulterait d'un simple principe pragmatique général prévoyant la possibilité et le devoir de contribuer à une question en cours, que dans une conversation ou un débat le droit de contribuer soit effectif et équitable. Avec comme résultat que des conflits peuvent éclater et que se met en place dans les domaines les plus sensibles un ensemble de normes collectives visant à assurer une certaine égalité dans l'interlocution. Des formes de réglementation des tours contributionnels sont ainsi mises en place [1], qui peuvent être complètement formalisées dans des domaines comme le domaine juridique ou politique. Les interactants disposent de quelques repères, comme des termes lexicaux cadrant la conversation : il s'agit de formes lexicalisées associées soit à la réalisation d'un commen-

1. On peut citer par exemple le fait que cela soit la défense qui dans un procès ait le dernier mot, cette norme tenant compte du fait qu'aux termes des débats, l'intervention finale a le pouvoir de laisser les participants dans une perspective attentionnelle particulière.

Expressions	Source	Métalangage [2]	Notes / Remarques	Maxime de Grice associée	Valeur ratification-nelle (+ ou - forte)	Impact prosodique (oui/non)
donc euh	CNTRL - asso. BOI_M1_09.txt	Revenir sur	On s'égare du sujet	Relation	Moyenne	Hausser le ton ?
bon ça fait vingt-deux minutes euh treize euh qu'on a commencé sur ce sujet, est-ce qu'on ne passerait pas à autre chose	CNTRL - asso. BOI_M1_09.txt	Commenter	Tentative de clore un sujet	Quantité	Forte	Impatience ?
c'est pas ce que je voulais dire euh et donc euh ouais mais non mais oui mais faut qu'on voit le repas quoi	CNTRL - parents / enfant. FETE_LEC_07.txt	Modifier	Rectification	Manière	Assez forte	Excuse ? Agacé ? Blasé ?
c'est tout ce qu'il a dit	Entendu dans des conversations	Commenter	Rapporter	Quantité	Moyenne	Déçu ? Etonné ? Agacé ? Ravi ? Désolé ?

TABLE 1 – Expressions méta-linguistiques (contributionnelles et discursives).

taire méta-contributionnel (par exemple : « vous oubliez de dire que... ») soit à sa description par un tiers (par exemple : X a rétorqué que). Une tentative de fournir un premier relevé et à en esquisser la typologie a été faite, en séparant notamment les formes qui sont relatives à des contraintes « gricéennes » (comme la contrainte de complétude) et celles qui n'en relèvent pas. Les expressions méta-contributionnelles servent à « recadrer » la conversation, à expliciter pourquoi tel locuteur souhaite attirer l'attention sur ce point en particulier, souligner quelque chose, ne pas laisser de place à l'ambiguïté. Celui qui emploie ce type d'expressions explique ce qu'il est en train de faire tout en le disant. Tout locuteur peut également employer des expressions méta-discursives, soit ce qui est dit à propos de ce qui est dit, c'est-à-dire faire un commentaire implicite mais « inclus » dans l'énoncé. La prosodie peut accentuer l'énoncé.

Lors d'une discussion, chaque participant tente de faciliter la compréhension de l'interlocuteur, et d'aller vers un consensus (Clark et Schaefer, 1989; Roulet, 1987). Les échanges sont soumis à des contraintes contributionnelles qui s'avèrent souvent très proches des contraintes contributionnelles classiques comme la maxime de quantité, de qualité ou de pertinence – *complétude* (pour reprendre le terme de Portugues (2011)) - à ceci près que l'enjeu de la contribution n'est pas informationnel mais est de définir ce qui doit être pris en compte et de la façon dont cela doit être pris en compte. De ce fait, l'interlocuteur s'appuie sur l'aspect sémantique de l'énoncé du locuteur (qu'est-ce qui est dit ? - presque mot pour mot -) ainsi que sur son aspect pragmatique (qu'est-ce qui est dit dans ce qui n'est pas explicitement dit ?) pour interpréter le tout et répondre/réagir en conséquence.

3 Méta-communication sur la ratification

Nous présentons Table 1 quelques exemples relevés dans des corpus audio.

2. indiquant que le locuteur veut revenir sur quelque chose, ou modifier quelque chose, ou commenter

4 La prosodie entre méta-communication et ratification

Il a souvent été supposé que l'orientation argumentative de ce qui est dit serait prévisible de son contenu sémantique et donc que les études de langue axées sur l'argumentation pourrait se concentrer exclusivement sur ce contenu afin de comprendre la dimension linguistique de mécanismes argumentatifs.

Nous avons au contraire plaidé pour la nécessité d'admettre que, parce que les contours prosodiques sont essentiels à la compréhension de « ce qui est dit à propos de ce qui est dit », mais aussi parce que ces commentaires au sujet de ce qui est dit ont souvent la capacité de modifier ultimement le contenu de « ce qui est dit », l'orientation de l'argumentation des énoncés ne peut jamais être prédite sans la considérer en détail. Nous avons tenté d'illustrer cette réalité en examinant le rôle des contraintes prosodiques dans la détermination de l'orientation argumentative des énoncés, y compris (et parfois réduite à) les signes linguistiques tels que enfin ou quelques ou vas-y ou oui. De ce fait, nous pourrons observer la nature de l' « information » fournie par les commentaires prosodiques associés à diverses utilisations de ces signes linguistiques en français. L'étude la plus complète - bien qu'en cours d'analyse donc non exhaustive et de semblant brut - présentée ici porte donc sur la diversité des formes prosodiques associées à la réalisation d'un oui, et à l'interprétation méta-argumentative à laquelle chacune d'entre elles donne lieu. Elle s'appuie sur un travail réalisé à partir des corpus d'emplois de oui d'un projet de recherche (2013-2015) et est illustrée par des exemples. L'intérêt est de montrer que le ton qu'on emploie est un commentaire sur ce qui est dit. Les résultats permettent aussi de constater que le oui peut être convaincu ou non, et qu'il est parfois plus une marque de politesse que de ratification, ou associé à des formes variées de réticence ou de réserve, et enfin qu'il peut même parfaitement vouloir dire non. La prosodie peut marquer une forme d'interaction complexe. On peut entendre dans les « oui » des jeux polyphoniques indiquant la position du locuteur, de l'interlocuteur, et la position du locuteur à prendre par rapport à l'interlocuteur.

Par la prosodie, le locuteur se donne les moyens de faire partager sa croyance et de ce fait il introduit un comportement chez son interlocuteur. Celui-ci va devoir prendre en compte la façon dont lui a été transmis le message pour ajouter un sens à ce qui a été dit.

Il existe des multiples façons de dire enfin, quelques, oui, ou vas-y, et qui ne peuvent être comprises que par la prosodie de l'énoncé. L'interlocuteur se basera sur cette prosodie pour enchainer (humour, ironie, obéissance, fausse approbation, etc.). La prosodie enrichit les modèles portant sur les contraintes du dire, car il y a une dimension polyphonique qui permet d'interpréter implicitement un énoncé. Cela donne un type d'orientation argumentative encore jamais étudié à ce jour.

5 Ratification, prosodie et théorisation des conversations

Si la prosodie a un rôle déterminant dans la compréhension d'un énoncé, c'est parce qu'elle permet à l'interlocuteur d'avoir des repères sur la suite (l'enchainement) des énoncés à fournir. En fait, lorsque nous parlons, nous ne nous basons pas uniquement sur les mots employés (leur sens) par notre interlocuteur pour lui répondre, nous rebondissons également (voire parfois uniquement – sous-entendu, implicite) sur la prosodie. On peut ainsi constater le décalage plus ou moins grand entre « ce qui est dit » et ce « qui est pensé » (dire oui quand on pense non, ça s'entend !). En observant l'enchainement des énoncés en analyse conversationnelle, nous avons pu constater qu'il existe une stratégie argumentative qui consiste à enchainer sur la façon dont le message a été transmis et non pas sur le contenu. Il est facile de trouver dans des discussions quotidiennes des cas où quelqu'un dit un « oui » qui veut dire « non » (pas convaincu/explicite) et que l'interlocuteur réponde « oh bah si tu le prends comme ça. . . ». L'interlocuteur peut toujours prétendre ne pas avoir dit X (après tout, il a bien dit le mot « oui »).

5.1 Construction de l'échange

La construction des échanges est gérée au fur et à mesure de la production par les co-participants, qui effectuent des choix dans l'instant de l'interaction (Skrovec, 2010). L'organisation informationnelle et l'organisation topicale, présentées par E. Roulet, L. Filliettaz, et A. Grobet s'avèrent assez pertinentes dans la construction des conversa-

tions (Kuyumcuyan, 2001). L'étude de l'organisation topicale (ou thématique) vise à rendre compte des faits de continuité et de progression du discours. Ils présentent également la dimension référentielle, expliquant que « parler c'est agir sur autrui ».

5.2 La parole comme demande de prise en compte de quelque chose

Ce mécanisme, que les psychologues nomment attention contrôlée, peut se traduire par le fait que très souvent parler revient à attirer l'attention de quelqu'un sur quelque chose en lui demandant de le prendre en compte, sachant par ailleurs que cela ne peut pas être fait de façon neutre et que l'on ne peut donc pas attirer l'attention de quelqu'un sur quelque chose sans lui indiquer d'une façon ou d'une autre comment il doit être pris en compte, ce que les psychologues appellent cette fois « référencement social ».

5.3 Elaboration d'un champ attentionnel partagé

Si la parole peut-être vue comme une demande de prise en compte de quelque chose, c'est parce qu'on ne peut pas attirer l'attention de quelqu'un sur quelque chose sans que cette personne ne comprenne pourquoi. Donc, que cela ne peut pas être fait de façon neutre et que l'on ne peut pas attirer l'attention de quelqu'un sur quelque chose sans lui indiquer d'une façon ou d'une autre comment il doit être pris en compte. A partir de là, il est possible de poser que les échanges conversationnels reposent sur l'existence d'un champ attentionnel partagé et sur un principe de présomption de contribution.

Conclusion

La question de l'interface sémantique/pragmatique est presque toujours posée au niveau des énoncés, or la question se pose en réalité au niveau d'un continuum contributions/séquences contributionnelles/conversation, dès lors en particulier qu'il y a bien marquage linguistique (prosodique) de la gestion pragmatique des séquences ratificationnelles et marquage prosodique de l'orientation argumentative des contributions. Les contraintes prosodiques sont des contraintes linguistiques qui ne relèvent pas d'une théorie de l'implicite et sont porteuses d'informations méta-contributionnelles et conversationnelles : on ne peut donc ni identifier la sémantique au niveau de la phrase non-intonée, ni, quand l'on prend en compte la phrase/séquence intonée, ignorer qu'un segment peut être associé à une prosodie qui concerne l'ensemble d'une contribution ou encore une séquence ratificationnelle.

La figure 1 récapitule le déroulement d'une séquence contributionnelle.

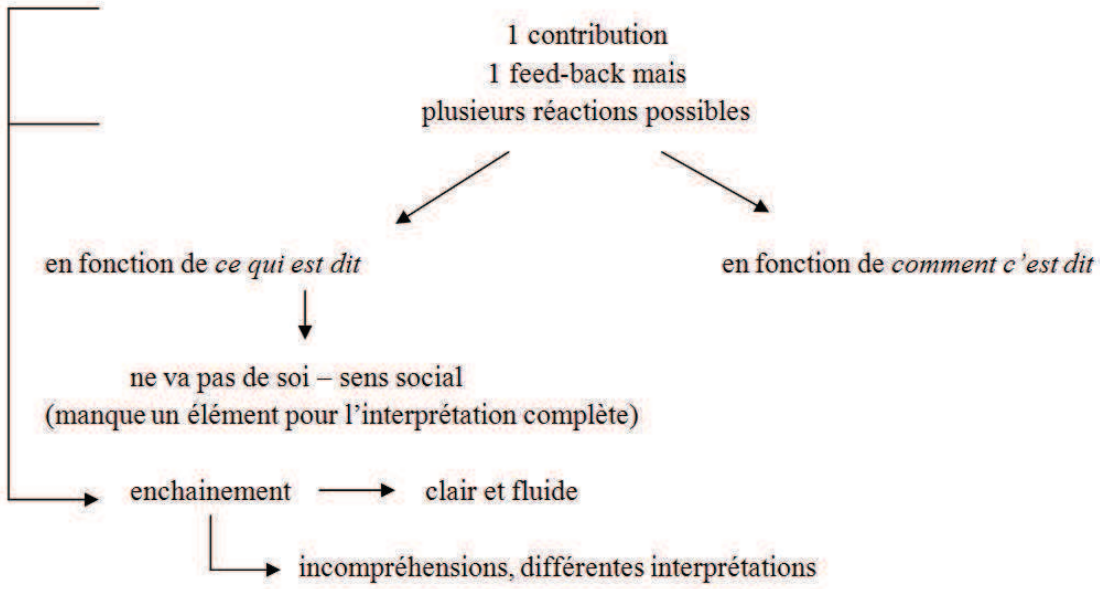


FIGURE 1 – Schéma récapitulatif du déroulement d'une séquence contributionnelle, prenant en compte la pragmatique, la sémantique, et l'analyse conversationnelle

References

J-C. Anscombre et O. Ducrot. 1976. L'argumentation dans la langue. *Langages*, (42) :5–27.

A. Auchlin et A-C. Simon. 2004. Gabarits prosodiques, empathie (s) et attitudes. *Cahiers de l'Institut de linguistique de Louvain-CILL*, 30(1) :181–206.

J. Authier-Revuz. 2004. La représentation du discours autre : un champ multiplement hétérogène. *Le discours rapporté dans tous ses états*, pages 35–53.

A.O. Barry. 2002. Les bases théoriques en analyse du discours. *Documents de la Chaire MCD*, 159.

J. Borderieux. 2013. *La construction textuelle du brevet d'invention : analyse et théorisation de la strate contributionnelle*. Ph.D. thesis, Université d'Orléans.

H. Clark et E. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2) :259–294.

M-M. de Gaulmyn. 1987. Reformulation et planification métadiscursives. J. Cosnier et C. Kerbrat-Orecchioni, editors, *Décrire la conversation*, pages 167–198. Presses Universitaires de Lyon.

H. P. Grice. 1975. Logic and conversation. P. Cole et J. L. Morgan, editors, *Syntax and Semantics : Vol. 3 : Speech Acts*, pages 41–58. Academic Press, San Diego, CA.

A. Kuyumcuyan. 2001. Lecture de Roulet, E. and Filliettaz, L. and Grobet, A. avec la collaboration de Burger, M. Un modèle et un instrument d'analyse de l'organisation du discours. *Cahiers de praxématique*, volume 37, pages 175–178. Presse Universitaire de La Méditerranée.

F. Nemo. 1999. The pragmatics of signs, the semantics of relevance, and the semantic/pragmatic interface. Ken Turner, editor, *The Semantics/Pragmatics Interface From Different Points of View*, pages 1–343. Elsevier.

F. Nemo. 2007. The pragmatics of common ground : From common knowledge to shared attention and social referencing. *Lexical Markers of Common Grounds. Amsterdam : Elsevier*, pages 143–158.

Y. Portugues. 2011. *Contraintes pragmatiques de complétude et linguistique des contributions en théorie du texte et de l'organisation textuelle : élaboration d'une heuristique appliquée au roman de formation*. Ph.D. thesis, Université d'Orléans.

P-Y. Raccah. 2011. Racines lexicales de l'argumentation : la cristallisation des points de vue dans les mots. *Verbum (Presses Universitaires de Nancy)*, 1(32) :119–141.

J. Rey-Debove. 1997. Le métalangage : étude linguistique du discours sur le langage. *Le Robert, Paris*.

E. Roulet. 1987. Complétude interactive et connecteurs reformulatifs. *Cahiers de linguistique française*, 8(111-140).

G-E. Sarfati. 2012. *Eléments d'analyse du discours*. Armand Colin.

M. Skrovec. 2010. *Répétitions : entre syntaxe en temps réel et rhétorique ordinaire*. Ph.D. thesis, Aix Marseille 1.

A. Steuckardt et A. Niklas-Salminen. 2005. Les marqueurs de glose. *Langues et langage*.

Prediction of Upcoming Words and Individual Differences in L2 Sentence Processing: an Eye-tracking Study

Verónica García-Castro
University of York/ University of Costa Rica
Department of Education
University of York, Heslington, York, YO10 5DD, UK
vgc505@york.ac.uk

Abstract

The ability to predict upcoming material can contribute in language interaction since language users may communicate faster when knowing what language material is coming (Kutas *et al.*, 2011). Studies have shown that word prediction is possible in adult monolinguals (Altmann & Kamide, 1999; Borovsky *et al.*, 2012) and in adult second language speakers (Kaan, 2014; Martin *et al.,* 2013). However, when it comes to second language prediction, whether L2 speakers predict upcoming material as L1 speakers still remains unclear, and whether individual differences have an effect on L2 predicting processes remains unexplored. The present work aims to find out to what extent L2 language users are able to predict upcoming words, and if the individual differences of phonological working memory, language aptitude, and vocabulary size have an effect on L2 prediction of upcoming words. The study's methodology is similar to the one adopted by Altmann & Kamide (1999).

Key Words: L2 word prediction, L2 sentence processing, eye-tracking, individual differences, phonological working memory.

1 Introduction

The ability to predict upcoming material can contribute in language interaction since language users may communicate faster when knowing what language material is coming. Kutas *et al* (2011) have mentioned that a potential benefit of prediction "is that it may allow a listener or reader to produce an overt response more quickly, without waiting for the material itself to become available" (Kutas *et al.*, 2011, p.190). Hence, language users may not need to receive all the input in order to communicate rapidly. If the predicted material is accurate, the speed of processing and communication will definitely increase. Nevertheless, when the upcoming material does

not match the prediction, reanalysis and reprocessing costs are likely to take place. The mismatch can be "used to adjust future predictions and minimize the chance of future errors (Jaeger & Snider, 2013)" (as cited in Kaan, 2014, p.257). In sentence processing, prediction can be an element of success or difficulty (Mehravari *et al*, 2015) where semantic, morpho-syntactic, and lexical aspects of the words yet to appear may be pre-activated (Federmeier, 2007, p.492). Pre-activation can contribute to the prediction of different specific aspects of the upcoming words (Federmeier, 2007) where language users may actively predict the word forms, semantics, morphology, and syntax of upcoming material (Fine *et al* (2013); Levy (2008), as cited in Mehravari *et al*, 2015). Therefore, language users may predict some, or all, of the aspects of the words yet to come and this can enhance their communication.

Native speakers use their lexical, syntactical, and semantic knowledge about a lexical item to predict upcoming material (Kaan *et al*, 2010), and different studies have shown some of the possible predictive mechanisms in adult native speakers. For instance Altmann & Kamide (1999) in a visual-world eye-tracking study have found that in monolingual sentence processing, it is possible to predict upcoming material when identifying the verb, preceding the direct object, when hearing sentences like: "The boy will *eat* the cake." In their study they presented the input as auditory material where participants' eyes´ movements were recorded, while looking at visual scenes, to determine their predicting processing. Participants listened to the auditory input while being presented with the visual scenes and they had to determine if the auditory input matched the scenes. One of the main findings of the study is that predictions in adults can occur when hearing and identifying the verb preceding the direct object in a sentence. In another eye-tracking study on L1 prediction, Borovsky *et al* (2012) discovered that adult native speakers make fewer predictions when their vocabulary size is smaller. In their study, participants' eye-movements were recorded while looking at visual scenes when listening to sentences such as "The pirate hides the treasure," and they had to click on the picture that matched the sentence. In the study, participants also took offline tests to estimate their vocabulary size. They analysed the anticipatory fixations and their relationship with age and vocabulary size. Their findings suggest that vocabulary size has an effect on anticipatory processing in adults, and thus, in their predictive processing. In another eye-tracking study, Kukona *et al* (2011) tested prediction of upcoming words in two different experiments. In the first experiment they used active sentences such as "Toby arrests the crook," and in the second experiment they used passive sentences like "Toby was arrested by the policeman." Their findings

suggest that local thematic priming can be relevant in word prediction, and that strong thematic relations can have strong effects on activation of upcoming material. Recently, Chow *et al* (2016) have been arguing that predictive mechanisms may also be related to memory retrieval. In sum, studies on L1 word prediction have found that there are diverse processes and mechanisms that drive the prediction of upcoming material.

When it comes to second language prediction, whether second language speakers predict upcoming material as L1 speakers still remains unclear. For instance, some studies have shown that second language speakers do not predict the upcoming materials as native speakers do (Kaan, 2014; Martin *et al.*, 2013). Some other studies have found that L2 language users may present native-like predicting processes (Hopp, 2013; Dissias *et al.*, 2013, as cited in Kaan, 2014). Even though second language speakers may have all the information necessary for prediction, their predictive process seems to be dissimilar from those of native speakers. To illustrate, Kaan (2014) has argued that the processing differences between second language speakers and native speakers is due to factors such as frequency information, where native speakers have received more quantitative and qualitative input than non-native speakers. Another factor is the competing information in the bilingual mental lexicon. It is known that both languages are activated during the parsing of either where L2 speakers may show non-native predictive patterns due to their lack of suppression of irrelevant candidates while making predictions (Kaan, 2016, p, 1).Therefore, second language users activate more information, when making predictions, and this can influence their predictive processing. In an eye-tracking study Grüter *et al* (2012) have found that the online predictive mechanisms in L2 grammatical gender diverge between native and non-native speakers. They tested, through the participants' eye movements, whether the gender-marking of the determiner would contribute to the prediction and interpretation of the following noun (p, 203). Their results show that native speakers were faster when identifying and looking at the target picture than non-native speakers; therefore, they seemed to make faster predictions. In an ERP study Martin *et al* (2013) tested whether second language speakers predicted to the same extent as first language speakers. They hypothesized that L2 comprehenders' prediction of upcoming words is slower than that of L1 comprehenders. Participants had to predict the final noun phrase in sentences with two different conditions: expected and unexpected endings. The N400 amplitudes found revealed that L2 comprehenders predict to a weaker extent than L1 speakers. In sum, studies on L2 prediction of upcoming material have found that there are differences between L1 and L2 predictive processes; however, more research is needed in order to have a deeper

understanding of the differences, if any, of those processes.

Up to this point, aspects on what drives the L1 and L2 predictive mechanisms have been discussed; however, cognitive differences among individuals have not been mentioned. In second language processing, the study of individual differences can contribute to our understanding of "how general cognitive skills and domain-specific skills jointly determine behavior" (Roberts & Meyer, 2012, p.3). Individual differences can be found in almost all cognitive activities (Eysenck & Kane, 2015, p.427); thus, when it comes to language processing, individuals may differ in the abilities they use for such processing. It is still unclear if individual differences have an effect or not, or to what extent, on the prediction of upcoming words.

An individual difference that has been previously research is phonological working memory (PWM). It is a crucial language learning device that assists the acquisition of novel phonological forms in first and second language learning (Baddeley, 2003; Baddeley *et al*., 1998). Research has shown that there is an association between PWM ability and L2 vocabulary acquisition (Speciale *et al.*, 2004), and that L1 phonological processing abilities facilitate L2 learning of unfamiliar phonology (Abreu & Gathercole, 2012). Thus, if PWM is crucial in language learning, would it have an effect on the prediction of upcoming words? Would participants with a higher PWM predict faster than participants with a lower PWM? Another individual difference researched in language processing is vocabulary size. Borovsky *et al.* (2012) have found that vocabulary size has an effect on L1 prediction, but it is still unclear if it has an effect on L2 prediction of upcoming words. Hence, it is necessary to include individual differences as possible factors in language prediction to obtain a better understanding of both L1 and L2 prediction of upcoming words and their underlying processes.

2 Present Work

The present work aims to find out whether L2 language users are able to predict upcoming words according to their subcategories and if this prediction differs from that of L1 speakers. By directly comparing subjects, in a within-subjects design, the study aims to find out whether or not the individual differences of PWM and vocabulary size have an effect on the predictive processes of L2 speakers. In this study, it is hypothesized that

1. The subcategories of verbs and nouns influence prediction, where verbs generate more prediction than nouns
2. L1 phonological short term memory facilitate prediction and the speed of processing of upcoming L2 words.
3. Larger vocabularies facilitate speed of processing in prediction of upcoming L2 words.

3 Methodology

The methodology of the study is similar to the one adopted by Altmann & Kamide (1999). A visual-world eye-tracking study will determine, through the participants' eye movements, if they are able to predict the upcoming material, while looking at visual scenes, before listening to the aural input. Part of the evidence of prediction processing in adults has been taken from "eye movements in response to language while viewing a visual scene" (Borovsky *et al.*, 2012, p. 418), which highlights the validity of the method in predictive processing.

The study takes into account prediction of nine nouns and nine verbs and their subcategories. Before taking the eye-tracking task, participants will take the battery of offline individual differences tests, then, they will be presented with written stimuli on a computer screen. They will read twelve different sentences per target word, where all the sentences are semantically and grammatically correct. After reading the stimuli, participants will take the visual eye-tracking task.

English-like nonwords will be used as the target nouns and verbs to predict in order to guarantee that participants have not had previous exposure to the target words. The nonwords were extensively piloted with thirty English native speakers and with thirty Spanish native speakers with an advanced proficiency level of English as a second language. The piloting was to make sure that all nonwords were equally guessable among both language users. In addition, the nonwords were piloted with ten English native speakers for phonotactic validity. Only those nonwords that were pronounced near-identically, among the native speakers, were taken into account for the study.

3.1 Offline Tests

One of the most effective tests to measure PWM is a nonwords repetition test (NWR). The NWR performance relies on the capacity to perceive, store, recall and reproduce phonological sequences (Juff & Harrinton, 2011), and it can give a " purer assessment of phonological storage quality than serial recall measures using lexical stimuli as memory items" (Gathercole 2006, p.520). Previous studies have successfully used NWR tests as a measure of PWM (O'brien *et al.*, 2006; Speciale *et al.*, 2004; Cheung, 1996; Gathercole, 1995); therefore, for the present study, a NWR test in the L1 (Spanish) will be used.

A vocabulary size test (Nation, 2012) will be used to account for vocabulary size, and a verbal fluency task (Rommers *et al*, 2015) for lexical availability.

3.2 Participants

The participants for the study will be 25 English native speakers studying at a university in the United Kingdom and 25 Spanish native speakers, with an advanced level of English

as a Second Language, studying at a university in the United Kingdom.

4 Results & Conclusions

The results of the study will contribute to the understanding of prediction of upcoming words in L2, if prediction is hindered or enhanced by word type and its subcategories, and how individual differences may have an effect on predictive processing. Even though it is known how relevant individual differences are in language processing, there is a lack of studies on L2 prediction of upcoming words that take them into account. Therefore, the study comes to fill a theoretical gap and to potentially bring more understanding in L2 prediction processes.

References

Altmaan, G., & Kamide, Y. (1999). Incremental Interpretation at Verbs: Restricting the Domain of Subsequent Reference. *Cognition* 73, 247-264. Retrieved from www.elsevier.com/locate/cognit

Baddeley, A.D., Gathercole, S.E. & Papagno, C. (1998). The phonological loop as a language learning device, *Psychological Review,* 105(1), 158-173. Retrieved from http://ovidsp.tx.ovid.com/sp3.22.1b/ovidweb.cgi?&S=PADNFPLNONDDANHGNCHKKCFBEGCMAA00&Link+Set=S.sh.18.19.22.25%7c7%7csl_10

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36, 189-208. doi.org/10.1016/S0021-9924(03)00019-4

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417–436. doi: 10.1016/j.jecp.2012.01.005

Chow, W., *et al* (2016). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition, and Neuroscience*, 31(5), 617-627. doi: 10.1080/23273798.2016.1160135

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition*, *35*, 353–387. doi: 10.1017/S0272263112000915

Eysenck, Michael., & Keane, Mark. (2015). *Cognitive psychology: a Student's Handbook*. (Revised 6th Ed). Abingdon, Oxon: Psychology Press.

Federmeier, K.D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4), 491-505. doi: 10.1111/j.1469-8986.2007.00531.x

Gathercole, S. (2006). Complexities and constraints in nonword repetition and word learning. *Applied Psycholinguistics*, *27*, 599-613. doi.org/10.1017/S014271640606053X

Grüter, T. *et al* (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research,* 28(2) 191–215. doi:10.1177/0267658312437990

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. doi: 10.1177/0267658312461803

Juffs, Alan., & Harrington, Michael. (2011). Aspects of working memory in L2 learning, *Language Teaching*, 44(2), 137-166. doi:10.1017/S0261444810000509

Kaan, Edith. (2016). Susceptibility to interference: underlying mechanisms, and implications for prediction. *Bilingualism: Language and Cognition,* 19, 1-2. doi:10.1017/S1366728916000894

Kaan, Edith. (2014). Predictive Sentence Processing in L2 and L1. *Linguistic Approaches to Bilingualism* 4(2), 257–282. doi: 10.1075/lab.4.2.05kaa

Kukona, A., *et al*. (2011). The time course of anticipatory constraint integration. *Cognition,* 119, 23-42. doi:10.1016/j.cognition.2010.12.002.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). *A look around at what lies ahead: Prediction and predictability in language processing*. In M. Bar (Ed.), Predictions in the brain: Using our past to generate a future (pp. 190-207). Oxford University Press.

Martin, Clara., et *al*. (2013). Bilinguals Reading in their Second Language do not Predict Upcoming Words as Native Readers do. *Journal of Memory and Language,* 69(4), 574-588. doi.org/10.1016/j.jml.2013.08.001

Roberts, Leah., & Meyer, Antje. (2012). Individual Differences in Second Language Learning: Introduction. *Language Learning*, 62(2), 1-4. doi: 10.1111/j.1467-9922.2012.00703.x

L’interface organisation linguistique / organisation poétique à la lumière de la théorie des actes de langage

STÉPHANE DUCHATELEZ,
Université de Toulon - Laboratoire Babel EA 2649 - École doctorale 509
40 bis avenue de Suffren, 75015, Paris
stephane.duchatelez@gmail.com

Résumé

La présente communication[1] propose une approche pragmatique d’un aspect fondamental de la communication poétique que sont les « effets poétiques ». La lecture d’un poème en vers libres de Michaux nous permet d’isoler trois types de phénomènes étroitement corrélés à la production d’effets poétiques : la présence d’une séquence énumérative dépourvue de thème-titre, l’observation d’organisations phonique, métrique et pseudo-strophique, et enfin un certain nombre d’effets sémantiques reposant principalement sur l’anaphore rhétorique, constitutive du poème. Le recours à la théorie des actes de langage nous permet de formuler l’hypothèse suivante : il y a effet(s) poétique(s) lorsque des effets perlocutoires sont produits sans que ces derniers reposent sur un acte illocutoire finalisé.

Mots clés : effets poétiques, parallélismes poétiques, actes de langage.

1 Introduction

Si l’on souscrit à l’idée selon laquelle il n’existe pas, à proprement parler, de « langage poétique », l’on doit admettre que, dans une large mesure, un poème reçoit de la part du lecteur un traitement interprétatif de type linguistique. Selon cette conception, le texte poétique « appartient au système énonciatif de la langue » (Hamburger, [1977] 1986 : 208). Ce postulat a d’ailleurs pu être vérifié par Monte (2002 ; 2003), qui, à partir du corpus des textes de Jaccottet, a démontré la pertinence d'une approche énonciative des textes poétiques.

Comment un texte donné accède-t-il cependant au statut de production poétique ? S’inspirant de Jakobson (1963 : 209-248), de nombreux poéticiens (Aroui, 1996 : 118-125 ; Dominicy, 2011 ; Gouvard, 2009 : 15-20 ; Ruwet, 1975 : 307-351) font l’hypothèse d’une seconde organisation, laquelle, sans s’y réduire, serait fortement corrélée à la présence de parallélismes. Tous les parallélismes ne relevant pas de l’organisation poétique, Dominicy (2011 : 64-65) a proposé des contraintes d’éligibilité au statut de parallélisme poétique. L’idée commune à ces auteurs est que les parallélismes poétiques produisent des effets irréductibles à l’interprétation linguistique. Nous partirons donc de l’hypothèse que les effets poétiques résultent d’un traitement alternatif au traitement sémantico-référentiel habituel[2].

Dans la présente communication[3], nous limiterons l’étude des effets poétiques à la présence de tels parallélismes. La démarche que nous adopterons consistera à interroger les modalités d’émergence de l’effet poétique à l’aune de la théorie des actes de langage[4]. Nous commencerons par préciser la version de la théorie des actes de langage sur laquelle reposent nos hypothèses. Puis,

[1] Cette communication s’inspire d’un travail actuellement en cours, dans le cadre d’une thèse de doctorat en Poétique, menée dans le département de Sciences du langage de l’université de Toulon, sous la direction de Michèle Monte. Le but de cette recherche est de conduire une réflexion sur les modalités d’apparition de l’effet poétique, en travaillant sur un corpus de poèmes de Henri Michaux et Jacques Roubaud, et en privilégiant une approche pragmatique et cognitive.

[2] Pour une approche « pertinentiste » de la question, voir par exemple Sperber & Wilson ([1986] 1989 : 326-336), Constable (1998) et Pilkington (2000).

[3] Je remercie Michèle Monte de sa relecture d’une version préliminaire de cette communication.

[4] Pour une application de la théorie des actes de langage aux textes poétiques, voir notamment Monte (2002 : 309-334), et surtout Dominicy (2009), qui en souligne la pertinence.

dans le but d'asseoir le débat sur une base empirique, nous proposerons une lecture d'un poème en vers libres du recueil *La vie dans les plis*, de Henri Michaux, en centrant l'analyse sur les phénomènes associés à la production d'effets poétiques. Pour terminer, nous tenterons d'interpréter ces phénomènes en corrélation avec la théorie des actes de langage.

2 Cadre théorique

Dans ce qui suit, nous nous appuierons sur une version cognitive de la théorie des actes de langage, et plus précisément l'approche propositionnelle qui a été développée dans Kissine 2007[5]. L'auteur y propose « un modèle psychologiquement plausible de la manière dont nous attribuons des forces illocutoires à nos énoncés » (p. 9). Y sont reprises et discutées les principales catégories commentées depuis Austin 1970.

Le niveau le plus élémentaire est composé des actes phonétique/graphique et phatique. Le premier repose sur la production de phonèmes/graphèmes relevant du système phonétique/graphique de la langue utilisée, tandis que le second consiste à doter le message d'« une signification et [d']une structure syntaxique dans la langue de communication » (Kissine, p. 92). Mais à la différence d'Austin, la composante rhétique y est incluse dans le niveau locutoire. Kissine considère en effet que l'acte locutoire renvoie à un « niveau de sens propositionnel » (p. 98), lequel permet d'exprimer un état épistémique tel qu'une croyance, un désir ou une intention (p. 185). Enfin, le niveau illocutoire consiste à doter le niveau locutoire d'une force assertive, directive, ou commissive (Dominicy, 2009 : 41). Dans ce qui suit, nous limiterons la discussion aux seuls énoncés assertifs. Kissine définit l'acte illocutoire assertif, où la proposition *p* est assertée, comme le fait, par l'énonciation de *p*, de donner à l'interlocuteur des raisons de croire que *p*. Ainsi, arrivant de l'extérieur un jour de pluie, l'énonciation de la phrase « il pleut » permettra au locuteur d'accomplir l'acte de donner à un interlocuteur des raisons de croire qu'*il pleut*. L'accomplissement de l'acte illocutoire entraîne alors la réussite ou l'échec de l'acte perlocutoire, lequel consiste, pour les actes assertifs, à amener l'interlocuteur à adopter la croyance que *p*.

Pour résumer, les niveaux phonétiques et phatiques servent de support au niveau locutoire, lequel est indispensable à la réalisation de l'acte illocutoire. L'effet perlocutoire, quant à lui, pourra être causé par un acte illocutoire, sans cependant en être constitué.

Les textes poétiques sont-ils régis, au même titre que les autres productions, par l'architecture que constituent les différents niveaux qui viennent d'être présentés ? L'hypothèse que nous souhaitons examiner ici est celle de l'irréductibilité des effets poétiques à ce modèle.

3 Le poème "Emplie de" de Michaux

Le poème « Emplie de », qui constitue la première des neuf pièces en vers libres du volume *Apparitions*, nous servira ici de matériau afin d'évaluer la robustesse de notre hypothèse :

EMPLIE DE

Emplie de moi
Emplie de toi.
Emplie des voiles sans fin de vouloirs obscurs.
Emplie de plis.
Emplie de nuit.
Emplie des plis indéfinis, des plis de ma vigie.
Emplie de pluie.
Emplie de bris, de débris, de monceaux de débris.
De cris aussi, surtout de cris.
Emplie d'asphyxie.
Trombe lente.
(*La vie dans les plis*, p. 74)

Au plan textuel, la dominante discursive du texte peut être ramenée à une séquence énumérative, qui consiste ici en une dérivation en série du constituant prépositionnel. Conformément au schéma prototypique de la séquence descriptive, dont l'énumération est une variante, l'on peut considérer que le dernier vers du poème relève de l'opération de reformulation, considérée par Adam comme l'une des quatre procédures descriptives (Adam, 1997 : 85-89).

Trois phénomènes distinguent cependant le poème du prototype énumératif : l'absence de thème-titre d'une part, sur lequel repose en principe toute séquence énumérative (Adam, 1997 : 85) ; la présence d'une organisation pseudo-strophique, qui dote la séquence énumérative de plusieurs sous-séquences ; enfin, la réduplication systématique par anaphore rhétorique du constituant participial « Emplie de », qui, nous le verrons, n'est pas sans poser problème au plan sémantique.

[5] Je remercie Marc Dominicy d'avoir porté à ma connaissance le travail de Kissine.

3.1 Un thème-titre absent

Quelles hypothèses peut-on faire sur le thème-titre ? Si l'emploi du féminin exclut l'hypothèse du pronom sujet masculin de première personne, l'idée d'un substantif métonymique du sujet-locuteur n'est cependant pas à écarter.

On relève de très nombreuses assonances en [i] dans l'ensemble du poème : 24 au total, auxquelles on peut ajouter les 2 semi-voyelles [ɥi] des vers 5 et 7. Le poème comptant au maximum 74 syllabes, cela représente un ratio de plus 32 % de syllabes contenant les phonèmes [i] ou [ɥi], soit près d'une syllabe sur trois. Il semble donc intéressant d'orienter la recherche vers un thème-titre contenant le phonème [i]. D'autre part, le quatrième vers – « Emplie de plis » – contient l'un des deux substantifs du titre du recueil. La syllabe [pli] y forme un parallélisme phonique et métrique [ã**pli** / də**pli**] qui invite à réévaluer en termes sémantiques la réduplication du participe au début de neuf des onze vers du poème. Il n'est pas indifférent en effet que le signifiant du titre soit phonétiquement représenté dans l'anaphore rhétorique. Le participe « emplie » ne serait donc pas sans évoquer le substantif « plis ». Le titre du recueil lui-même, au moyen de parallélismes phonique et métrique qu'il instaure entre « vie » et « plis », concoure à rapprocher les deux termes.

L'ensemble de ces remarques nous paraissent suffisamment convaincantes pour formuler l'hypothèse que le syntagme *ma vie*, où le possessif renvoie au poète, constitue un thème-titre plausible du poème. La structure profonde de chacun des énoncés du poème serait donc constituée de la formule : *Ma vie est emplie de …*

Conformément aux hypothèses de Ruwet (1975 : 323, 329-330), l'organisation poétique, au moyen de parallélismes phoniques et métriques, accompagne ici un défaut de l'organisation discursive – l'absence de thème-titre –, et permet de formuler une hypothèse sur la nature sémantique du thème-titre. Dans ce qui suit, nous admettrons la validité d'une telle hypothèse.

3.2 Une organisation pseudo-strophique

Si l'on observe l'ensemble du texte, on remarque une nette dominante des unités tétrasylalbiques (vers 1, 2, 4, 5, 7). A ces cinq vers, il convient de remarquer que le syntagme « Emplie de bris » du vers 8 prolonge en le reproduisant le mètre tétrasyllabique du vers précédent, que la superposition du zeugme et de la ponctuation contribuent à rendre d'autant plus saillant. De même, l'octosyllabe du vers 9 peut être segmenté en deux hémistiches tétrasyllabiques, comme le suggère la ponctuation qui accompagne la reformulation par épanorthose, ainsi que la « rime » interne entre « aussi » et « cris ». On aboutit donc à l'organisation métrique suivante : 4 / 4 / 11 (12 ?) / 4 / 4 /14 / 4 / 13 (4 / 3 / 6)/ 8 (4 + 4) / 5 / 3 (2 ?).

On constate une relative alternance entre vers isométriques tétrasyllabiques et vers hétérométriques. Cette dernière catégorie est composée d'un hendécasyllabe au vers 3 (bien qu'il soit envisageable de lire ce vers comme un alexandrin, en prononçant le e muet de « voiles »), d'un vers de 14 syllabes au vers 6, d'un vers de 13 syllabes au vers 8, d'un pentasyllabe au vers 10, et enfin du dernier vers de 3 syllabes (ou 2 syllabes si l'on ne prononce pas le *e* muet). On peut avec Aroui (2000 : e4) considérer que les vers hétérométriques, en instaurant une métrique contrastive, concourent à renforcer les équivalences interstrophiques. L'hypothèse d'une organisation pseudo-strophique, où chaque groupe serait constitué d'un ou plusieurs tétrasyllabes suivis d'au moins un vers hétérométrique, paraît donc plausible.

Un certain nombre de phénomènes vont par ailleurs venir se superposer à ce matériau pseudo-strophique, et contribuer soit à le renforcer, soit à le nuancer. La présence par assonance, dans les syntagmes prépositionnels des trois premiers vers, de la semi-voyelle [wa] – [mwa], [twa], [vwal], [vulwaʁ] –, tandis que les compléments des vers 4 à 10 se caractérisent par une distribution systématique des phonèmes [i] ou [ɥi], incite en effet au regroupement des trois premiers vers. La double occurrence de « plis », aux vers 4 et 6, instaure une circularité qui confère au sixième vers un effet de clôture. Le fait que l'anaphore rhétorique ne soit qu'implicite au vers 9, ainsi que l'emploi de l'adverbe « aussi » favorisent son regroupement avec les lignes 7 et 8. Au contraire, la reprise, au vers 10, de l'anaphore rhétorique – éludée au vers précédent –, ainsi que le volume syllabique du substantif « asphyxie », concourent à isoler le vers du groupe précédent. Enfin, les nasales [õ] et [ã] du dernier vers, corrélées à l'absence du phonème [i], contribuent à l'instauration d'une sonorité contrastive avec l'ensemble du poème, à laquelle s'ajoute l'absence de structure participiale. Il ressort de ces remarques l'hypothèse organisationnelle suivante : 4 4 11 (12) - 4 4 14 - 4 13 8 - 5 - 3 (2).

3.3 Un poème allégorique

Munis d'hypothèses sur le thème-titre ainsi que sur un éventuel séquençage du texte, nous allons à présent proposer des éléments d'interprétation du poème.

Un montage allégorique

Associés au thème-titre supposé *la vie*, le parallélisme à travers lequel apparaissent les deux pronoms disjoints « moi » et « toi », ainsi que l'absence de ponctuation entre les deux premiers vers, suggèrent une lecture métonymique des pronoms, les référents physiques y étant interprétés à partir du thème duel des personnes.

Succédant au couple *moi-toi*, que formalise la structure actancielle *locuteur-adressé(e)*, le terme psychologique abstrait « vouloirs » du troisième vers s'interprète aisément à partir de la thématique du couple et du désir amoureux, entraînant la possibilité d'une lecture figurale du sens de « voiles ». L'élaboration d'un motif abstrait dans ces premiers vers se heurte cependant au caractère concret ou phénoménal d'un grand nombre de syntagmes prépositionnels dans la suite du poème, par ailleurs susceptibles d'activer le sens spatial du participe « Emplie de ». Deux parcours interprétatifs semblent alors possibles. L'on peut choisir de poursuivre la lecture abstraite du poème, en privilégiant l'interprétation métaphorique des constituants nominaux. Le syntagme *emplir* y acquerra alors un sens attributif. Mais l'on peut également proposer une lecture concrète de chacun des vers, lecture susceptible de s'appuyer nous le verrons sur l'existence d'une logique thématique et événementielle. Aucun des deux niveaux ne prévalant sur l'autre, le poème relèvera d'un dispositif général allégorique[6]. Ici, l'absence de thème-titre, et le caractère sous-entendu du plan interprétatif abstrait en font une allégorie implicite (Bonhomme, 1998 : 72-74).

Sens métaphorique

La lecture figurale de « voiles » autorise une comparaison du désir avec l'image de voiles tendues par les vents marins. Mais associé à « vouloirs », le substantif « voiles » peut également s'interpréter comme la propriété que possède le désir d'être dissimulé, caché, voilé. Il s'agira ici d'une interprétation par hypallage[7], le syntagme « voiles (...) de vouloirs » s'analysant à partir de celui de *désirs voilés*. Hypothèse que semble confirmer la présence de l'adjectif « obscurs », qui partage avec *voilés* l'idée d'obstacle à l'accessibilité.

Des remarques d'ordre phonologique permettent par ailleurs d'affiner la lecture des trois premiers vers. L'assonance en [wa] s'accompagne d'une configuration consonantique composée essentiellement de consonnes antérieures labiales (« em**p**lie », « **m**oi », « **p**lis », « **v**oiles », « **f**in », **v**ouloirs ») ou apicales (« **t**oi », « **d**e », « **d**es », « emp**l**ie », « voi**l**es », « vou**l**oirs », « **s**ans »). Harmonie que vient rompre l'adjectif final « obscurs », dans lequel figurent deux consonnes postérieures, d'abord la vélaire [k], suivie de l'uvulaire [ʁ]. Ces phénomènes phonologiques peuvent être croisés avec des remarques d'ordre interprétatif. L'analyse de « voiles » en *voilés*, que nous venons de proposer, est compatible avec le stéréotype classique du jeu de la dissimulation amoureuse. L'adjectif « obscurs » qui termine l'ensemble vient rompre cette monotonie, la dissimulation volontaire y étant remplacée par l'expression d'une incapacité à accéder aux motifs du désir, difficulté qui peut concerner le désir de l'autre comme le sien propre, et qui coïncide avec la survenue des consonnes postérieures. L'évolution du groupe formé par les trois premiers vers consiste donc, selon notre interprétation, en le déroulement du motif de la réciprocité amoureuse, vers sa problématisation abrupte et finale, que semble d'ailleurs confirmer le caractère dysphorique de la suite du poème. Le lecteur se trouve ainsi orienté vers un stéréotype qui s'avère être une fausse piste interprétative, et reproblématisé à la toute fin du troisième vers.

L'interprétation du lexème « plis » pourra être rapprochée d'une expression figurée telle que *les plis du cœur*, que l'on peut paraphraser comme « [l]a partie la plus intime, la plus secrète du cœur » (*CNRTL*, entrée « pli »). Ce choix, qui consiste à exprimer l'image d'une zone peu accessible (« plis ») et négative (« nuit ») de l'être, a l'avantage de s'articuler avec le thème du désir

[6] Le texte de Michaux s'inspire à plusieurs égards du poème « Mouvement » de Rimbaud (Arthur Rimbaud, *Œuvres complètes*, Pléiade, 2009), dont il reprend le motif allégorique du « Vaisseau », ainsi que certains éléments lexicaux (« Trombe ») et thématiques (celui du couple). Toutefois, notre propos étant centré sur l'organisation poétique, la question de la relation à l'intertexte rimbaldien se limitera à cette note.

[7] En réalité, les termes « voiles » et « vouloirs » échangeant leurs fonctions grammaticales, il s'agira d'une double hypallage (voir Bernard Dupriez, *Gradus. Les procédés littéraires*, Éditions 10/18, 1984, pp. 235-236).

amoureux de la première strophe, et en particulier avec le SN « vouloirs obscurs » du troisième vers. Au vers 5, le lexème « nuit » peut recevoir une interprétation symbolique par activation du sème afférent /inconnu/ ou /inaccessible/. Si le domaine sémantique[8] sélectionné est //moral//, l'interprétation se rapprochera du sens du lexème *mauvais*. Enfin, l'occurrence de « vigie », en associant les sèmes /surveillance/ et /dangerosité/, concoure à activer l'idée de tension dans la dynamique amoureuse, et rend plausible l'hypothèse d'une allusion à des rapports conflictuels.

Au vers 7, le substantif « pluie », succédant par paronomase *in praesentia* à celui de « plis », semble en être le produit (Fromilhague, 1995 : 24). Conventionnellement, le terme peut activer le sème afférent /dysphorie/, comme en témoigne l'expression par hypallage « un temps triste », et se trouve donc susceptible de décrire les revers de la dynamique amoureuse. De même, le lexème « bris » peut dans le contexte adéquat suggérer l'idée d'altération de la relation amoureuse[9], ce qui renforce l'hypothèse interprétative du groupe précédent.

Les « cris » peuvent par métonymie exprimer une crise au sein du couple, et l' « asphyxie » est une métaphore conventionnelle de la difficulté de vivre dans un environnement perçu comme excessivement clos ou contraignant. On sait que la rédaction du volume *Apparitions* suit de peu la tuberculose de Marie-Louise Michaux (Martin, 2003 : 412-416). Or, il est attesté que la maladie de Marie-Louise a été vécue par Michaux comme une contrainte qui « impos[e] [...] comme par une sorte de fatalité, le rythme implacable de l'autre » (Martin, 2003 : 414).

On peut également proposer une interprétation abstraite du dernier vers, en considérant que les traits inhérents /puissance/, mais surtout afférents /violence/ et /danger/ peuvent être réinterprétés dans le cadre de la thématique amoureuse. Mais à partir de quel référent de l'univers abstrait réinterpréter le référent concret correspondant au substantif « trombe » ? Si l'on considère que la fonction de reformulation du dernier vers entre en relation prédicative d'équivalence avec le thème-titre (Adam, 1990 : 181), on pourra supposer que la vie du locuteur est présentée comme équivalant à une « Trombe lente ». Mais il est également possible d'envisager que le référent abstrait de « Trombe » renvoie à la relation amoureuse elle-même, et que le constituant final reformule le sens abstrait du corps du texte.

Sens concret

Comme le suggère le caractère concret ou phénoménal des substantifs du poème, une lecture non métaphorique du texte est également envisageable. Nous allons voir qu'il est possible de s'y appuyer sur la présence d'une logique thématique et événementielle.

La double occurrence, aux vers 4 et 6, du substantif « plis », peut s'analyser de deux façons. On peut d'abord considérer que les deux occurrences renvoient à un même référent. Dans cette hypothèse, il est plus plausible de concevoir une relation causale entre les termes « plis » et « nuit » : le pli, parce qu'il engendrerait de l'ombre, serait porteur de nuit, le terme étant ici synonyme d'obscurité. A quel type de référent les expressions peuvent-elles être associées ? La présence du lexème « voiles » dans la première strophe active le sème générique //maritime//. Or, le même domaine sémantique apparaît au vers 6 avec le substantif « vigie ». Ce dernier peut s'interpréter comme le poste d'observation, la personne occupant ce poste, ou l'activité d'observer elle-même. Le terme « plis » du vers 6, et rétrospectivement celui du vers 4, peuvent dans ce contexte évoquer les plis des voiles visibles depuis le poste d'observation.

L'hypothèse d'un référent unique pour les deux occurrences de « plis » se heurte cependant au rôle joué par la répétition. La reprise par épanorthose paraîtrait en effet plus plausible si les deux occurrences étaient immédiatement successives. L'insertion du vers 5 entre les deux occurrences affaiblit en effet nettement l'hypothèse d'une reformulation. S'il est peu contestable que l'occurrence de « plis » du vers 6 renvoie aux plis des voiles d'un navire, il est tout à fait possible de concevoir que celle du vers 4, ainsi que l'occurrence de « nuit », décrivent des propriétés des vagues observables depuis un navire.

Si l'on poursuit la thématique du domaine maritime, l'on peut imaginer que la « pluie » au vers 7 renvoie à des phénomènes d'intempéries, et que les « bris » et « débris » au vers 8 en constituent les dommages matériels.

[8] Sur la notion de sème spécifique, afférent et générique, voir Rastier 1987.

[9] Cet emploi est par exemple attesté chez Huysmans : « (...) songeant actuellement, devant son feu, au **bris** de ce ménage qu'il avait aidé, par ses bons conseils, à s'unir, il jeta une nouvelle brassée de bois, dans la cheminée, et il repartit à toute volée dans ses rêves. » (Joris-Karl Huysmans, *A Rebours*, éd. Marc Fumaroli, Paris, Gallimard, coll. « Folio », 1977, p. 163, nous surlignons).

Les vers 9 et 10 introduisent l'idée d'un danger imminent, dont le vers 10 représenterait le point culminant. L'on peut considérer que les « cris » et l'« asphyxie » résultent d'intempéries en mer. Une hypothèse alternative, inspirée d'éléments biographiques, permet d'interpréter les « cris » comme les symptômes de la maladie de Marie-Louise, et l'« asphyxie » comme sa phase la plus aiguë, les lexèmes y étant distribués par gradation sur une échelle exprimant la gravité.

Le dernier vers, dont la structure profonde est entièrement substantive, a un statut particulier en raison de sa fonction de reformulation et de clôture. Le substantif « Trombe » entretient de multiples liens isotopiques avec l'ensemble du poème. Le fait qu'il existe des trombes marines rend le terme compatible avec l'isotopie maritime qui parcourt l'ensemble du poème. Le terme entre également en écho avec le substantif « vigie » du vers 6, par activation des traits afférents /dangerosité/ et /violence/. Enfin, il reprend l'idée d'altération de l'environnement, ce qui est conforme au champ lexical de la dévastation – « bris », « débris » – du vers 8. Le dernier vers permet donc de nommer l'événement organisateur du niveau concret de l'allégorie.

L'interprétation que nous avons proposée du poème à partir du sens concret des syntagmes prépositionnels fournit donc les éléments d'une trame événementielle. Le scénario ainsi proposé ressemble à celui d'une catastrophe maritime :

- vers 4-6) : attente du danger (« plis », « nuit », « vigie »)
- vers 7-10 : intempéries (« pluie »), dégâts matériels (« bris », « débris ») et humains (« cris », asphyxie)
- (v. 11) : reformulation (« Trombe lente »)

Rapportée à la sphère intime du locuteur-énonciateur, la stratégie allégorique vise donc, selon notre interprétation, à dresser un portrait sombre de l'univers du couple à partir du scénario d'une catastrophe maritime.

3.4 Interpréter l'anaphore rhétorique

Si elle met bien en lumière l'existence d'une dimension thématique et événementielle du propos, notre analyse accorde cependant peu de place à un certain nombre de phénomènes qui concourent à perturber cette logique événementielle. Par principe, une séquence énumérative repose sur des relations de juxtaposition entre les différents éléments qui la composent. Le caractère énumératif du texte a donc pour conséquence d'atténuer les effets narratifs éventuels découlant de la logique événementielle sous-jacente. Les liens entre les différents éléments de la séquence se trouvent donc de ce fait affaiblis. L'emploi de l'adverbe « aussi » au vers 9 suggère en effet que les liens entre les lignes obéissent a une logique de juxtaposition, ce qui confère une dimension cumulative plutôt que chronologique ou causale au propos. Le statut particulier du vers 10, isolé des autres pseudo-strophes, ne fait qu'accentuer ce phénomène. Enfin, la réduplication systématique du segment « Emplie de » a pour effet additionnel de souligner la présence des référents associés à chacun des syntagmes prépositionnels, au détriment d'une représentation globale. Ce surcroît de focalisation est ainsi responsable d'une tendance à l'autonomisation de chacune des lignes.

A ces considérations d'ordre général, il convient d'ajouter quelques remarques de nature sémantique. Aux vers 7 et 8, les termes « pluie » et « bris » (« débris ») peuvent être chacun associés à des référents distincts. L'impossibilité de réduire la lecture du niveau concret à un graphe sémantique unique doit cependant être analysée en tenant compte du trait /intensif/ du verbe *emplir.* Dans le *Nouveau Petit Robert*[10], l'emploi y est qualifié de VIEILLI ou LITTÉRAIRE, et le verbe *remplir* est suggéré : « Rendre plein (un réceptacle), utiliser entièrement (un espace disponible) ». Or, deux référents concrets distincts ne sauraient emplir chacun et simultanément le même espace. La même remarque s'applique à des référents de nature phénoménale, comme les « cris » et l' « asphyxie ». Il y a donc conflit entre le sémantisme du verbe et ses différentes réduplications dans des énoncés portant chacun sur un référent distinct. La reformulation, au vers 9, de « aussi » en « surtout », en reproduisant le sens intensif d'emplir, pourrait même suggérer que les différentes représentations s'annulent les unes les autres.

Le choix de certains termes induit des effets de sens qui concourent également à virtualiser l'inscription référentielle du propos. L'interprétation du niveau concret des lignes 4 et 5 consistait à dire que les « plis » pouvaient être ceux des vagues aperçues depuis la vigie du navire. Il faut cependant tenir compte du fait que le lexème *pli* est un concept abstrait. Hors de toute interprétation métaphorique, et lorsqu'il renvoie à un référent singulier, il est toujours associé, explicitement ou non, à un second substantif : *les plis de la robe, les plis du drap*, etc. Le fait de présenter

[10] *Le Nouveau Petit Robert de la langue française*, Dictionnaires Le Robert, 2009 : entrée « Remplir ».

le terme indépendamment de toute notion complémentaire oriente le travail interprétatif vers un sens générique, et affaiblit d'autant la recherche sémantique d'un référent individuel. Ces remarques entrent d'ailleurs en écho avec l'une des acceptions de l'adjectif « indéfinis » (vers 6), qui consiste à exprimer ce « [q]ui n'est pas clairement défini, qui, n'étant pas spécifié, demeure vague » (*CNRTL*, entrée « indéfini »). La même analyse s'applique à l'occurrence de « nuit », dont le sens concret renvoie habituellement à une notion unitaire, de nature météorologique et temporelle – attendre la nuit –, parfois itérative – travailler de/la nuit –, difficilement compatible avec l'emploi massif induit par la structure partitive. Les difficultés d'activer le sens concret du lexème paraissent ici aussi favoriser une interprétation générique, qui renverrait à une idée de ce que serait l'essence de la nuit, mais dont l'inscription dans le niveau concret du dispositif allégorique resterait toute virtuelle. Ces phénomènes, associés aux effets produits par l'anaphore rhétorique, vont contribuer à les renforcer, et favoriser à la fois l'autonomisation et la virtualisation des représentations qui résultent de la lecture de chacune des lignes.

4 Les effets poétiques et la théorie des actes de langage

Dans ce poème, l'absence de thème-titre fournit l'élément premier du caractère indéfini du texte. Ces difficultés pour le lecteur d'articuler le propos à partir d'un thème organisateur précis a pour corollaire la présence de phénomènes phoniques et pseudo-strophiques qui contribuent à orienter les hypothèses interprétatives. Nous avons en effet suivi un certain nombre de pistes, dont une hypothèse sur le référent du thème-titre, laquelle découlait de l'observation et de la mise en corrélation de parallélismes phoniques et métriques dans le corps du texte et dans le titre du recueil. La reconnaissance d'une organisation pseudo-strophique, elle-même reposant largement sur des phénomènes métriques, a d'autre part facilité la reconnaissance d'une trame événementielle pour le niveau concret. Le niveau infra-locutoire, dont relèvent les phénomènes phoniques et métriques, n'est normalement pas perçu en tant que tel, et ne subsiste habituellement pas au-delà de quelques secondes dans la mémoire de travail de l'interprétant (Tsur, 1996 : 57). Les textes poétiques semblent cependant avoir pour stratégie d'en prolonger le maintien en mémoire, et surtout d'y focaliser l'attention de l'interprétant (Schaeffer, 2015 : 105-112). Nous avons également montré que les effets résultant de l'anaphore rhétorique, lesquels relèvent principalement du niveau locutoire, conduisaient l'interprétant à entretenir simultanément plusieurs croyances incompatibles. Enfin, nous venons de voir qu'à cela s'ajoutaient divers phénomènes locaux qui contribuaient à virtualiser le propos.

L'approche que nous proposons consiste à soutenir que l'ensemble de ces phénomènes concourent à affaiblir le traitement illocutoire des énoncés. L'absence de thème-titre empêche quiconque d'entretenir la croyance que celui-ci serait *la vie* avec la même force que s'il était explicitement formulé. De même, la difficulté, qui découle de la présence de l'anaphore rhétorique, d'entretenir simultanément plusieurs croyances entraîne l'affaiblissement de la portée illocutoire de chaque proposition. Le traitement pragmatique du sens concret[11] des propositions comprend donc bien un niveau locutoire, mais sans que celles-ci se voient dotées d'une visée illocutoire finalisée. Si l'on reprend la définition de l'illocutoire telle qu'exposée au début de cet article, le locuteur du poème – le poète – ne serait pas supposé y donner des raisons de croire qu'*une chose serait emplie de pluie et* – simultanément – *serait emplie de débris*. Par corollaire, il faut supposer que l'interprétant n'adopte pas la croyance qu'*une chose serait emplie de pluie et* – simultanément – *serait emplie de débris*. Cependant, nier la présence d'effets interprétatifs comme la formation de représentations mentales découlant de la lecture reviendrait à nier toute pertinence à la modalité allégorique, pourtant constitutive du poème. Il est par conséquent indispensable de supposer l'existence d'effets perlocutoires.

En quoi consistent ces effets ? Dans la version de la théorie des actes de langage que nous avons privilégiée, le niveau locutoire est prépositionnel, et permet d'exprimer croyance, désir ou intention. L'effet perlocutoire va donc consister à favoriser dans l'esprit de l'interprétant la formation de différentes représentations d'un espace empli de référents divers, et correspondant aux syntagmes prépositionnels. Toutefois, privées de portée illocutoire finalisée, ces représentations ne sont pas converties en croyances par l'interprétant, ce qui lui permet notamment d'entretenir simultanément plusieurs représentations en prin-

[11] Ce problème ne se pose pas pour le niveau métaphorique, étant donné que le verbe *emplir* y possède un sens attributif et non locatif.

cipe paradoxales. On peut supposer qu'elles évoquent les croyances en question, au sens de Dominicy 2011, mais sans que l'interprétant les adopte pour autant.

De façon identique, certains phénomènes infralocutoires – phonétiques ou métriques – ont contribué à produire des effets interprétatifs, qu'ils soient propositionnels ou non – hypothèse sur le thème-titre, émergence d'une organisation pseudo-strophique ou événementielle –, sans que l'on puisse considérer ces derniers comme des croyances au sens fort du terme. L'hypothèse d'effets perlocutoire sans visée illocutoire paraît également plausible pour ces niveaux.

5 Conclusion

Les effets poétiques étudiés dans le cadre de cette communication consistent, d'après notre interpétation, en la production d'effets perlocutoires, sans que ces derniers reposent sur un acte illocutoire finalisé.

Cette définition de l'effet poétique est susceptible d'enrichir l'analyse des critères d'éligibilité au statut de parallélisme poétique proposée par Dominicy (2011 : 64-68) d'un volet pragmatique : seront éligibles les parallélismes produisant des effets perlocutoires ne reposant pas directement sur le niveau illocutoire.

Enfin, la définition de l'effet poétique que nous proposons ne se limite pas, en principe, aux seuls textes exhibant une organisation poétique. Aussi, les textes poétiques en prose, qui ne possèdent pas ou peu d'organisation poétique, devraient pouvoir être examinés à l'aune de nos hypothèses.

Références

Adam Jean-Michel, *Éléments de linguistique textuelle*, Mardaga, coll. « Philosophie et langage », 1990.

Adam Jean-Michel, *Les textes : types et prototypes*, Nathan, 1997.

Aroui Jean-Louis, « Nouvelles considérations sur les strophes », dans Dominicy et Michaux, éds., *Approches linguistiques de la poésie*, *Degrés*, n°104, 2000, pp. e1-e16.

Austin John Langshaw, *Quand dire, c'est faire*, Paris : Seuil, 1970.

Bonhomme Marc, *Les figures clés du discours*, Seuil, 1998.

Constable John, « The Character and Future of Rich Poetic Effects », in Shoichiro Sakurai, (ed.), *The View from Kyoto: Essays on Twentieth-Century Poetry*, Rinsen Books: Kyoto, 1998, pp. 89-108.

Dominicy Marc, « La théorie des actes de langage et la poésie », *L'Information Grammaticale*, n° 121, 2009, pp. 40-45.

Dominicy Marc, *Poétique de l'évocation*, Paris : Classiques Garnier, 2011.

Fromilhague Catherine, *Les figures de style*, Nathan, 1995.

Gouvard Jean-Michel, « Poésie, parallélisme et stéréotype dans l'œuvre d'Yves Bonnefoy », *L'Information Grammaticale*, n° 121, 2009, pp. 15-20.

Hamburger Käte, *Logique des genres littéraires* [1977], Paris : Seuil, 1986.

Jakobson Roman, *Essais de linguistique générale* [*I. Les fondations du langage*], Minuit, 1963.

Kissine Mikhail, *Contexte et force illocutoire. Vers une théorie cognitive des actes de langage*, Thèse de doctorat, Université libre de Bruxelles, 2007 [accessible en ligne sur le site « Di-fusion » de l'Université libre de Bruxelles].

Martin Jean-Pierre, *Henri Michaux*, Gallimard, coll. « Biographies », 2003.

Michaux Henri, *La vie dans les plis*, Gallimard, coll. « Poésie », 1972.

Monte Michèle, *Mesures et passages : Une approche énonciative de l'œuvre poétique de Philippe Jaccottet*, Paris : Honoré Champion, 2002.

Monte Michèle, « Essai de définition d'une énonciation lyrique. L'exemple de Philippe Jaccottet », *Poétique*, n° 134, 2003, pp. 159-181.

Pilkington Adrian, *Poetic Effects: A Relevance Theory Perspective*, Amsterdam/Philadelphie, John Benjamins, 2000.

Rastier François, *Sémantique interprétative*, Paris : PUF, 1987.

Ruwet Nicolas, « Parallélismes et déviation en poésie », dans Julia Kristeva *et al.*, éds, *Langue, discours, société. Pour Émile Benvéniste*, Paris, Seuil, 1975, pp. 307-351.

Schaeffer Jean-Marie, *L'expérience esthétique*, Gallimard, 2015.

Sperber Dan, Wilson Deidre, *La pertinence*, Minuit, [1986] 1989.

Tsur Reuven, « Rhyme and Cognitive Poetics », *Poetics Today*, vol. 17, n° 1, 1996 : pp. 55-87.

The Importance of Using Psycholinguistic Tools for CNL Evaluations

Nataly Jahchan
CLLE, University of Toulouse, Airbus Operations SAS
[nataly.jahchan@{airbus.com, univ-tlse2.fr}]

Abstract

Using psycholinguistic tools and evaluations has not been a common practice in the study of Controlled Natural Languages (CNLs). Human-Oriented controlled languages (languages destined to improve human comprehension of text) have mostly been the fruits of industrial needs in a human factors perspective. Increasing readability for human operators and decreasing text complexity in a human machine interaction context were the main concerns for industry. In this paper, we will show when and how these psycholinguistic evaluations have been used in the CNL domain, and the eventual shortcomings that we would like to focus and work on in order to improve the link between these two disciplines. We proposed the systematic use of more rigorous psycholinguistic tools to eliminate any form of bias in future evaluations, and a scale for evaluating the "naturality" of a CNL has been proposed.

Keywords: Psycholinguistic Evaluations, CNLs, Human-oriented CNLs, Naturality scale, Controlled Natural Languages

1 Introduction

The first CNLs had the aim of facilitating communication between humans. After World War I there was a need to have a common linguistic tool that the international community could use to communicate together (basic English 1930). After that, there was a growing need for CNLs in Industry. To name a few well-known ones: Caterpillar Fundamental English was used as a means of cutting costs on translation manuals for international human operators of Caterpillar machines. AECMA Simplified English was developed for maintenance manuals across different aircraft manufacturers. The International Civil Aviation Organization developed ICAO phraseology for air traffic control. Finally, the Airbus Controlled language was developed to enhance pilot comprehension of on-screen information in the cockpit.

Most CNL researchers agree that there are three main types of CNLs: Ones that improve comprehensibility, otherwise known as Human-oriented or Comprehension-oriented controlled languages, which are considered the origin of controlled languages. Translation-oriented CNLs are ones mostly used in natural language processing for automatic translation. And formal representation controlled languages that provide representation for formal logic sequences.

2 Definitions: The Many Faces of CNLs

It is important to provide definitions that encompass all the aspects of the various types of controlled languages, and by doing that exclude the languages that do not fall in the realm of CNLs (for instance: Languages that do not obey constitutive rules of base language, non-constructed languages that arise naturally like sublanguages, or languages that are not based on one language like Esperanto, and formal languages that are not intuitive enough to be understood by a native speaker of the language they are based on).

Kittredge (2003) provides a CNL definition that is somewhat comprehensibility-oriented as *"a restricted version of a natural language which has been engineered to meet a special purpose, most often that of writing technical documentation for non-native speakers of the document language. A typical CL uses a well-defined subset of a language's grammar and lexicon, but adds the terminology needed in a technical domain."* Whereas Fuchs and Schwitter (1995) define CNLs in a translation and formal representation oriented sense as *"a subset of natural language*

that can be accurately and efficiently processed by a computer, but is expressive enough to allow natural usage by non-specialists".

Kuhn (2014) on the other hand provides a comprehensive short definition of controlled language as *"... a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics, while preserving most of its natural properties."* He continues to say that CNLs are not necessarily proper subsets of the underlying natural language because there can be small deviations from natural grammar and semantics in addition to some unnatural elements like colors that are meant to increase readability. *"The subset relation is clearly too strict to cover a large part of the languages commonly called CNL."*

What we refer to here as CNL has been called many different names over the past: Controlled, processable, simplified, technical, basic, structured languages, guidelines, phraseologies etc. Kuhn (2014).

2.1 Controlled Natural Languages: Input and output

CNLs are constructed languages that must be based on one language; preserve most of the natural properties of the base language while being more restrictive. It is important however to differentiate between the input of a CNL which is its base natural language and its output which is not necessarily a very "natural" language. The word "natural" in the name controlled *natural* language is somewhat misleading because it refers to the input language and is not an accurate description of the resulting CNL. We could make the parallel in the field of Natural Language processing which also has the natural language as an input language in most cases. A CNL can vary in its dimension of naturalness on the PENS classification scheme (precision, expressiveness, naturalness, and simplicity, Kuhn (2014)) from N3 to N5, with N3 describing languages that have some natural and unnatural elements, but that are nevertheless understood by speakers of the language to a substantial degree; and N5 on the other end of the scale, describing languages that contain sentences with natural text flow.

3 Naturality Scale

We would like to propose a "Naturality scale" which is a work in progress at this stage, and on which CNLs would be placed on a continuum ranging from "Least naturalistic" or very coded to "Most naturalistic" or natural language in its theoretical state. In other words, the Naturality component could be roughly defined as the naturalness levels present in a language on a boundless continuum ranging from pure code to natural language. In this theory, natural language will always be theoretically unattainable ∞. Language is almost always to some extent controlled. Whether it is the written word or the spoken word, context, audience, aim, social decorum, even language rules, and many other outside factors force the user of the language to control to a certain degree what language he or she produces at a certain period in time. Therefore, language will not be divided into controlled and natural but should be placed on a naturality continuum with regards to all its aspects and the continuity of its gradations. This differs to the PENS classification scheme (but does not necessarily exclude it), because here we consider that the "naturality" aspect is the most significant dimension and from which all other dimensions should follow suit. PENS' aim is to describe and give qualifications of CNLs and not rate them, which would fit right along the naturality continuum.

From this we argue that the 4 dimensions that make up PENS (precision, expressiveness, naturalness, and simplicity) could be concatenated and placed onto one dimension of naturality. For depending on whether a language is naturalistic or not and where it should be placed on the naturality continuum, we would be able to extricate whether or not a language is precise (from many interpretations to extremely precise), expressive (from no quantification to able to express everything), or simple (virtually indescribable rules (NL) to described in one page).

Most importantly the classification of CNLs on the naturality continuum should be fluid because being subsets of natural language means that their application could hardly and fractionally be formalized in a clearly defined range. Additionally, CNLs as any language tend to evolve with time and with the need and application we have for them.

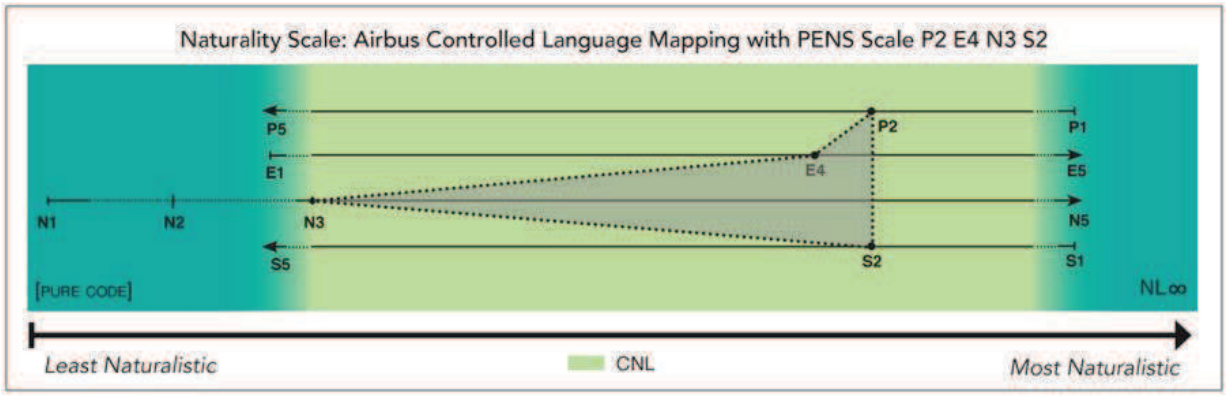


Figure 1. Naturality Scale: Airbus Controlled Language Mapping with PENS

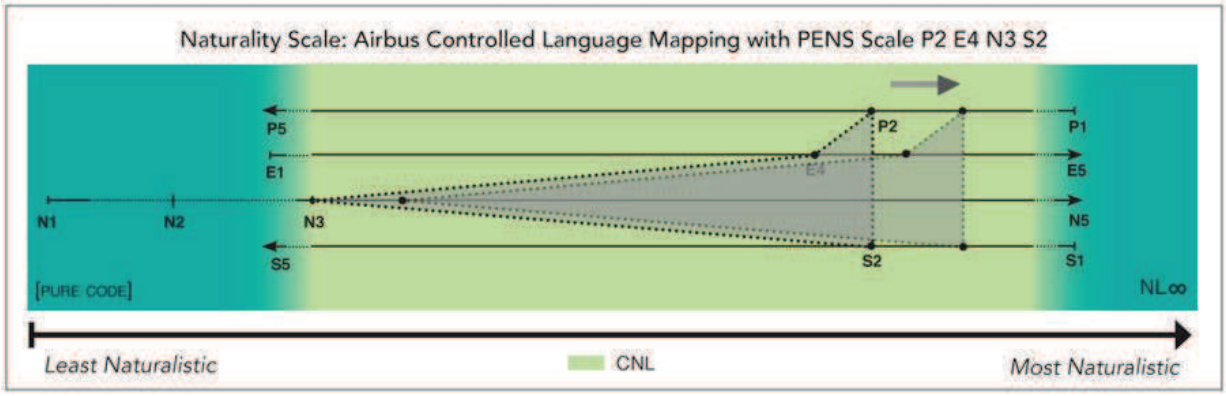


Figure 2. Naturality Scale: Airbus Controlled Language Mapping with PENS potential shift towards natural language

As we can see on the naturality scale **(Figure 1)**, we plotted the Airbus Controlled Language using what we assume this language to be classified on the PENS classification scheme P2 E4 N3 S2. The Precision, expressiveness, simplicity and naturalness are all plotted on the naturality continuum from least naturalistic [pure code] to natural language [NL ∞]. The Airbus Controlled Language forms the shape we see in the middle of the scale. What is interesting and novel about this representation is the fluidity with which a language can travel on the continuum. Considering the fluidity of languages, if a CNL becomes more or less naturalistic (as a result of an evaluation) and thus shifts on the continuum, the entire mapped CNL shape will shift accordingly since the foundation of this scale is the naturality continuum, the x-axis (example in **Figure 2**). Additionally, this scale also gives us a visual dimension of a CNL's naturality and could form grounds for comparison of different controlled languages that differ in their naturality levels and in their naturality evolution in time. Therefore, the Naturality scale is essentially a mapping of the PENS classification and criteria on a naturality based continuum. In other words, if a controlled language has become more natural as a result of psycholinguistic or other forms of experimentation (for example, if it was shown that there is a need to reduce the use of syntactical ellipses), it will shift on the naturality scale towards the most naturalistic side of the scale (right side), i.e. it becomes more natural. What this means is that when a language becomes more naturalistic it necessarily also shifts away from all its previous PENS dimensions. In this case **(Figure 1** and **2)** the new language becomes less simple to explain with traditional language rules (Simplicity dimension shifts from S2 to S1.5, the more natural a language is the less simple it is to Expressiveness explain). It will also be able to express more (dimension shifts from E4 to E4.5) etc. See **Figure 2** for a visual representation of this example.

3.1 Naturality Scale: Finding the right balance between natural and controlled

"Natural language being such a breeding ground for ambiguity, to communicate just one set of meanings while excluding many others is often impossible." (Crystal (1969) investigating English style) but it is also considered to *be "a universal tool of representation and of thought communication"* (Bisseret (1983)) and by others to represent the *"language of thought"* (Fodor (1975) that bears close resemblance to our surface language. *"In particular the syntax that governs the language of thought may be very similar or identical to that of external language. Studying syntax may therefore provide a window onto fundamental cognitive processes."* (Trevor A Harley (2013)).

Consequently, uncontrolled natural language is ambiguous and unsuitable for use in domains where ambiguity may be dangerous such as the aviation industry, but on the other hand, it represents an intricate part of our cognitive processes and its rules must not be excluded. Readability, text simplification, and text complexity research have focused on simplifying the language by making it less and less like natural language, and more like an unambiguous set of codes and regulations so that the resulting language veered away from the "natural" dimension. But to what extent is that simplification satisfactory and what are the limits at which it becomes counter-productive? When must natural language structures be respected?

4 Psycholinguistic Tools in the CNL Domain, an Overview

We believe the answer to that must lie in the systematic psycholinguistic evaluations of any established CNL and its various rules. *"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind. It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science."* William Thomson, Lord Kelvin.[1]

To this date CNL evaluations are not systematically enforced, and more than that they are very rarely put in place for human-oriented CNLs. There have been some evaluations of CNLs using NLP (natural language processing) tools in corpus linguistics based approaches such as the verification of requirements conformity (Condamines and Warnier (2014) or for text complexity Tuleshki Tanguy (2009), and machine translation O'brien and Roturier (2007), Aikawa et al. (2007) among others; There have also been evaluations based on Ontographs for formal representation languages Kuhn (2010)). But these evaluations fail to enlighten us on the effectiveness of these languages on the human cognitive processes of language comprehension, for instance by measuring reaction times and accuracy in performance.

The absence of empirical proof in the field has rarely (but not never) been criticized. Flesch (1944) criticizes Ogden (the creator of Basic English) for *"deliberately avoid[ing] the scientific approach and not [being] lucky enough to find the key to simplicity by accident"*. According to him, Linguists have criticized Basic English in an issue of the Saturday review of Literature for being *"a kind of quack based on a faulty analysis of the language process."* Nonetheless, Flesch (1944) concludes by saying that *"Basic English is the first attempt in the history of mankind to create a simplified language within a language […] and that simplified English is bound to come […] in a generation or two [...] and will be taken over by whatever system of simplified English we are going to adopt"*. Evidently, it is in fact the case. Hinson (1991) also criticized the absence of empirical proof: *"AECMA's Simplified English claims to be founded on readability research. It would be interesting to establish the nature, validity, and appropriateness of the research used. It would also be helpful to know of any research carried out on Simplified English manuals in use."*

To this effect, there have been some research in the mid-90's Shubert et al.(1996), Chervak et al. (1996), Chervak (1996), Eckert (1997), Stewart (1998) and again Temnikova (2012) that have attempted to acquire the much needed empirical evidence that speak to the added value of using controlled languages in certain corpora rather than their natural language counterpart.
A summary of these experiments will be shown in **Table 1** from Jahchan et. al (2016).

[1] http://uchicago.edu/~jagoldsm/Webpage/index.html

Author/year	Shubert et al. 1996	Chervak et al. 1996	Chervak 1996	Eckert 1997	Stewart 1998	Temnikova 2012
Native and non-native	Both	Both	Native	Non-native	Non-native	Both
Participants: natives	90 natives	157 natives	18 natives	0 natives	0 natives	22 natives
Participants: Non-natives	31 non-natives	18 non-natives	0 non-natives	148 non-natives	41 non-natives (21 different countries)	83 non-natives
Profession	Engineering students	AMT's	9 maintenance students and 9 experienced mechanics	Aviation maintenance students	Electronics technician students	All walks of life(because not testing SE, but CLCM)
Country	English speaking	English speaking	English speaking	Non-English speaking (Mexico)	English speaking	N/A (Online experiment)
Procedure	Reading comprehension, between subject	Reading comprehension, between subject	Performing maintenance, between subject	Reading comprehension, between subject	Reading comprehension, between subject	Reading comprehension, between subject
Tested for English comprehension	No	Yes (but not specifically for non-natives)	No	Yes	Yes	No (only self-evaluation and not used in analysis)
General SE Significance: doc type	Yes	Yes	No (means followed trend)	No (means followed trend)	No (means followed trend)*	No (means followed trend)
Significance SE comprehension: easy	No	No	N/A	N/A	N/A (only 1 workcard)	N/A
Significance SE comprehension: difficult	Yes	Yes	N/A	N/A	N/A (only 1 workcard)	N/A
Significance: time/SE	No**	No (will not adversely affect)	No	N/A	No	No
Significance: time/native speaker	N/A	Yes (normal)	N/A	N/A	No	N/A
Significance type of workcards	N/A (only easy/difficult was tested)	Yes (only certain workcards)	N/A	N/A	N/A	Yes (only certain sets of text)

*The study also concluded that the SE participants required higher mean English-reading ability to obtain a mean task card test score similar to the non-SE participants.

What is interesting, however, is that for procedure B (easier) the subjects reading **SE versions of that document took slightly **longer** than those reading the Non-SE versions.

Table 1. Jahchan et al. (2016). Comparative table summarizing most relevant results of different CL evaluations.

The complete analysis of the table could be found in the article it refers to. But what we will focus on here is the procedures that have been used to determine whether the controlled languages (AECMA SE and CLCM) significantly improved performance with regards to time and accuracy of responses over its natural language counterpart. None of these studies showed that the controlled language used improved the response times and only 2 out of 6 studies showed that Simplified English was more significant with regards to accuracy, and in those two studies there was a significant interaction with difficult tasks. The more difficult the task the more significant the Simplified English was. Therefore, it was task and document specific. The evaluations were somewhat inconclusive, but the controlled language was deemed good enough to be used since it did not adversely affect comprehension (and it is still being used across different aircraft manufacturers in maintenance manuals).

Reading comprehension was the procedure that has been used in the AECMA SE studies 1 to 5 in **Table 1**. That is, the maintenance students or the aircraft maintenance technicians had to read an aircraft maintenance procedure (either in AECMA SE CNL or pre-AECMA SE workcards, considered "natural language", even though it is hard to believe that the original technical writers did not control the text for ambiguities to a certain degree) and then reply to a multiple-choice questionnaire. The time it took participants to reply to these questions was recorded. Temnikova (2012) used a similar approach in an

online reading comprehension experiment where participants had to read emergency instructions in either the original "complex" text or in CLCM (Controlled Language for Crisis Management). The time for reading the text was limited. The results were evaluated using two evaluation metrics, percentage of correct answers and the time it took for participants to reply, which was not limited.

4.1 Psycholinguistic tools and the lack of proper evaluations

While these evaluations are a good effort, reading comprehension tasks do not accurately evaluate the real comprehension of a certain text, as the results will strongly rely on memory and skill. Additionally, with reading comprehension tasks we open ourselves to many uncontrolled biases such as the unlimited time that the participants have to answer after they have read a whole text with many details. In these evaluations, the texts were always about a maintenance procedure or an emergency task to be performed yet the participants did not perform the task but merely replied to questions about the task. In other words, we do not know whether the actions that are described in the text are accurately understood, whether they would have been correctly performed as such. We could only conjecture to the potential comprehension of a text that describes an action that the participants will not be performing. Therefore, these evaluations' shortcomings are due to the nature and assessment of the task itself. Proper psycholinguistic evaluations that accurately test human comprehension are an aspect that is missing in the human-oriented CNL domain.

We argue that the relative lack of psycholinguistic evaluations, barring the previous mentioned studies, is equivalent to rendering CNLs mere style guides or good authoring practices, and the reasons for adopting certain rules over others merely anecdotal.

Psycholinguistics uses psychological and neurobiological factors that enable us to study how the brain processes, comprehends, and acquires languages, etc. In short, it is the psychology of language. When we use psycholinguistic tools in CNL evaluations, we are merely proving linguistic hypotheses using psycholinguistic methods (behavioral tasks, eye tracking, Event Related Potentials). We are not learning about the function of the brain via models of psycholinguistics but rather, using psycholinguistic and psycho-cognitive methods to satisfy linguistic ends, in this case, the effectiveness of CNLs.

4.2 Psycholinguistic tools and a proposed protocol

The two disciplines must come together in a more effective manner, one that would reap the benefits of a tightly controlled psycholinguistic behavioral protocol evaluating reaction times and accuracy of comprehension in real-time participant performance. Such an experiment is currently under way. We are psycholinguistically testing the Airbus Controlled Language that pilots currently use in the cockpits to navigate and operate the planes against a more naturalistic (in syntax and lexicon) controlled language. Empirical results are being analyzed presently and will be the subject of a future publication.

5 Conclusion

This paper gives a brief overview of the current state at which CNLs stand in today's world. More particularly, it sheds light on the methods and evaluations that are used to assess the effectiveness of CNLs. It proposes a naturality scale that is a work in progress in order to have the possibility to plot any CNL on a scale that ranges from least to most naturalistic, as we argue that this is the most important dimension that characterizes a CNL and from which all other dimensions follow. We also propose an interpretation of the PENS scheme on this scale. Finally, we discuss the times that psycholinguistic tools were used in the human-oriented CNL domain, their shortcomings, and we proposed the systematic use of more rigorous psycholinguistic tools to eliminate any form of bias in future evaluations.

References

1. Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., & Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. Proceedings of the MT Summit XI, 1-7
2. Bisseret, A. (1983) Psychology for man computer cooperation in knowledge processing. In R.F.A. Masson (Ed.), IFIP 83, Information Processing 83.
3. Chervak, S. (1996). The Effects of Simplified English on the Performance of a Maintenance Procedure. Master's Thesis. State University of New York
4. Chervak, S., Drury, C. and Ouellette, J. (1996). Simplified English for Aircraft Workcards. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 40(5), pp.303-307

5. Condamines, A., & Warnier, M. (2014). Linguistic Analysis of Requirements of a Space Project and Their Conformity with the Recommendations Proposed by a Controlled Natural Language. In Controlled Natural Language (pp. 33-43). Springer International Publishing.
6. Crystal, D., & Davy, D. (1969). Investigating English Style
7. Eckert, D. (1997). The Use of Simplified English to Improve Task Comprehension For non-native English Speaking aviation maintenance technician students. Doctoral Dissertation, West Virginia University, WV
8. Flesch, R. (1944). How Basic is Basic English?. Harper's Magazine, 188(1126), 339-343
9. Fodor, J. A. (1975). The language of thought (Vol. 5). Harvard University Press.
10. Fuchs, N. E., & Schwitter, R. (1995). Specifying logic programs in controlled natural language. In Proceedings of CLNLP 95, 16 pages, Edinburgh
11. Harley, T. A. (2013). The psychology of language: From data to theory. Psychology Press.
12. Hinson, D. E. (1988). Simplified English—Is it really simple?. In Proceedings of the 38th International Technical Communication Conference
13. Jahchan, N., Condamines, A., & Cannesson, E. (2016, July). To What Extent Does Text Simplification Entail a More Optimized Comprehension in Human-Oriented CNLs?. In International Workshop on Controlled Natural Language(pp. 69-80). Springer International Publishing.
14. Kittredge, Richard I. 2003. Sublanguages and controlled languages. In Ruslan Mitkov, editor, The Oxford Handbook of Computational Linguistics, pages 430–447
15. Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. Computational Linguistics, 40(1), pp.121-170
16. O'Brien, S., & Roturier, J. (2007). How portable are controlled language rules? A comparison of two empirical MT studies. Proceedings of MT summit XI, 345-352
17. Shubert, K. Jan H. Spyridakis, Heat, S. (1995). The Comprehensibility of Simplified English in Procedures. Journal of Technical Writing and Communication, 25(4), pp. 347-369
18. Stewart, K. (1998). Effect of AECMA Simplified English On the Comprehension of Aircraft Maintenance Procedures By Non-native English Speakers, University of British Columbia
19. Tanguy, L., & Tulechki, N. (2009). Sentence complexity in French: A corpus-based approach. Proceedings of IIS (Recent Advances in Intelligent Information Systems), 131-145
20. Temnikova, I., (2012). Text Complexity and Text Simplification in the Crisis Management Domain. Ph.D. thesis, University of Wolverhampton

Dictionnaire électronique (DE) des noms simples issus de verbes Les noms issus des alternances *mp-* ou *f-*

Joro Ranaivoarison
Université d'Antananarivo
Centre Interdisciplinaire de Recherche Appliquée au Malgache
Madagascar
jororanaivo@yahoo.fr

Résumé

Cet article décrit la construction d'un dictionnaire électronique de noms issus de verbes du malgache (DEMA-NVS). Ces noms se composent de noms d'agent, de noms de profession, de noms de manière, de noms d'instrument, de noms d'action et de noms exprimant un état. Les structures morphologiques de ces derniers sont détaillées puis décrites à l'aide de transducteurs afin de construire une ressource destinée à des utilisations informatiques – un dictionnaire électronique. On discute dans cet article de la mise en œuvre du dictionnaire, du dictionnaire électronique lui-même et de son évaluation en rapport avec sa couverture lexicale.

Mots-clés : dictionnaire électronique, ressource linguistique, morphologie, malgache, nom

1 Introduction

Ce travail se situe à l'interface entre morphologie descriptive et traitement automatique des langues (TAL). Son objet est le malgache, une langue « peu dotée » en outils et ressources au sens de Berment (2004). Pour développer des outils de TAL qui rendent possible le traitement automatique de cette langue et permettre aux utilisateurs de disposer des moyens pour communiquer dans leur langue, il est nécessaire d'augmenter la couverture lexicale actuelle de celle-ci. En effet, il sera plus facile pour les développeurs d'applications de décider ou non de créer d'applications pratiques (correcteur grammaticale et/ou orthographique, outil d'aide à la traduction) pour le malgache si les ressources crées pour celui-ci ont une couverture correcte de ses lexiques, c'est-à-dire que tous les mots de la langue, du moins ceux se rattachant aux grandes catégories grammaticales (verbes, noms, adjectifs, adverbes, pronoms, etc.) sont insérés dans les ressources. L'objectif de cet article est de construire un dictionnaire électronique (DE) des « noms simples »[1]. Notre travail en cours porte sur 3 200 lemmes verbaux dont nous recensons des dérivés exprimant un état ou servant de noms d'agent, de profession, de manière, d'instrument ou d'action.

Le malgache est une langue agglutinante avec une riche morphologie, qu'il s'agisse de formes fléchies ou de formes dérivées. Dans cet article, une partie de la morphologie nominale est exposée. En effet, dans cette langue, il y a les mots qui sont eux-mêmes "noms" (N) comme *angady* "bêche, pelle", *trano* "maison", *penina* "stylo", *bara* « A. barre qui sépare les mesures en musique. B. Traverse, pièce mise en travers », *baby* « A. Épi de maïs sur la tige. B. Action de porter sur le dos. ». Ensuite, il y a les noms qui sont issus des alternances de l'élément temporel des verbes (V) avec *f-* ou *mp-* comme dans *mpijery* **N.** "spectateur", *fijery* **N.** "manière de regarder", *fijerena* **N.** "action de regarder" issus respectivement des verbes *mijery* **V. actif-statif** (act.-stat.) "regarder" et *ijerena* **V. circonstanciel** (circ.) "regarder". Enfin, il y a les noms qui sont issus des adjectifs (A) comme *hatsara* **N.** ou *fahatsara* **N.** « l'état de ce qui est bon, beau », *hatsarana* **N.** ou *fahatsarana* **N.** « la bonté, la beauté" issus de l'adjectif *tsara* **A.** "bon, qui a de bonnes qualités, beau ». Dans ce qui suit, seuls les noms issus des alternances du préfixe de temps avec *mp-* ou *f-*, qui sont des préfixes formatifs de nom, sont discutés. Ces noms sont issus de formes verbales comme *milalao* **V. act.-stat.** « jouer » dont dérivent *mpilalao* **N.** « joueur » (nom d'agent) et *filalao* **N.** « manière de jouer » (nom de manière) ; ou comme *anendrikendrehana* **V. circ.** « calomnier » dont dérive *fanendrikendrehana* **N.** « action de calomnier » (nom d'action). Toutes les fois que le terme "noms" est utilisé dans ce qui suit, il désigne les noms issus de cette formation.

Dans ce papier, les caractéristiques morphologiques des noms puis les méthodes utilisées (Gross, 1989) pour construire le dictionnaire sont présentées.

[1] Un dictionnaire électronique des verbes contenant 3 200 radicaux verbaux pouvant générer plus de 60 000 formes verbales a été déjà réalisé (Ranaivoarison *et al.,*2013, 2015a, 2016).

Par la suite est décrite la construction des graphes nécessaires au bon fonctionnement du dictionnaire avec Unitex, une plateforme de traitement de corpus écrits par dictionnaires et grammaires (cf. Paumier, 2016). Le dictionnaire électronique des noms issus de verbes simples (DEMA-NVS) et celui des paradigmes flexionnels des radicaux verbaux formant des noms simples (DEMA-NVSflx) sont ensuite présentés, ainsi que les résultats de leur évaluation.

2 Caractéristiques morphologiques des noms

Rajaona (1972, p. 642 - 645) présente les grandes lignes de la structure morphologique des noms issus des alternances du préfixe de temps avec *mp-* ou *f-*, préfixe formatif de noms, en malgache. Généralement, ces noms sont :

- soit des noms d'agent (Nag) ou de profession (Nprof),
- soit des noms de manière (Nman) ou d'état (Nét),
- soit des noms d'instrument (Ninst),
- soit des noms d'action (Nact).

Les noms d'agent et de profession sont à préfixe *mp-* se combinant avec les affixes de l'actif-statif[2] – les affixes de l'actif-statif sont : *i-, a-* ou une de ses variantes *an-, am-, ana-* apparaissant entre le préfixe de temps et le radical (cf. Rajaona, 1972, p. 454) – comme *mpijery* « celui qui regarde » analysé *mp-i-jery*, *mpandraharaha* « administrateur » analysé *mp-an-draharaha, mpamoha* « celui qui réveille, qui fait lever » analysé en *mp-am-oha* où *i-, an-, am-* sont des préfixes à valeur d'actif-statif. Les noms de manière, d'instrument et d'état sont à préfixe *f-* se combinant pareillement avec les affixes de l'actif-statif comme *fanafaingana* « manière d'accélérer » analysé *f-ana-faingana*, *famaky* « hache » analysé *f-am-aky*, *fihanjahanja* « l'état de ce qui est nu » analysé *f-i-hanjahanja* où *ana-, am-, i-* sont des préfixes de l'actif-statif. Enfin, les noms d'action se forment également sur *f-* avec des affixes à valeur de circonstanciel – les affixes à valeur de circonstanciel sont les affixes parasynthétiques du type *x-…-ana* où *x-* est un préfixe de l'actif-statif (cf. Rajaona, 1972, p. 159) – comme *fivoriana* « réunion, assemblée, séance » analysé *f-i-vori-ana, fihantsiana* « action de provoquer » analysé *f-i-hantsi-ana, fiverenana* « action de retourner » analysé en *f-i-veren-ana* où l'affixe parasynthétique *i-…-ana* est à valeur de circonstanciel.

Il s'ensuit que *mp-* est un préfixe formatif de noms d'agent et de profession[3] ; et, *f-* peut être :

- soit un préfixe formatif de noms de manière, d'instrument et d'état (quand il se combine avec les affixes de l'actif-statif)
- soit un préfixe formatif de noms d'action (quand il se combine avec les affixes du circonstanciel).

Pour aboutir à une description linguistique précise de chaque élément verbal pouvant former des noms, ces informations linguistiques[4] fournies par Siméon Rajaona (1972), en plus des informations sur les variations de formes des lemmes, sont codées et insérées dans le dictionnaire servant à une analyse morphologique claire et précise des noms de la langue.

3 Codification des noms simples

La morphologie à deux niveaux (Koskenniemi, 1983) a été largement utilisée pour traiter les langues agglutinantes telles que le finnois (Koskenniemi et Church, 1988), le turc (Oflazer, 1993) et même le malgache (Dalrymple *et al.,* 2006). Dans notre approche du traitement automatique du malgache, les méthodes analogues à celles utilisées pour le coréen (Nam, 1994 ; Nam et Paumier, 2014) ont été adoptées. Ces méthodes reposent sur des lexiques construits manuellement par des linguistes et ne sont pas à base de règles de calcul. Si les méthodes à base de calcul et/ou de statistiques ont l'avantage d'être économiques, les méthodes par dictionnaire sont précises et ont l'avantage d'être souples en ce qui concerne la maintenance et la mise à jour. Notre méthode de travail s'inscrit dans cette deuxième catégorie.

Elle se fonde sur les travaux de Gross (1989). La méthode se base sur une description explicite et détaillée de chaque mot de la langue. Rakotoalimanana (2000) mentionne cette approche. Sa description du malgache est explicite et claire et couvre tous les niveaux d'analyse (phonétique, morphologie, syntaxe, sémantique) et toutes les catégories grammaticales en allant dans les détails des découpages des affixes. Cependant, il ne mentionne que quelques exemples de variations morphologiques des mots, et ne vise pas une couverture lexicale substantielle. Par exemple, pour les verbes, son modèle ne prévoit pas d'indiquer pour chaque lemme verbal à quelle voix il peut apparaitre, ni quels affixes il prend parmi ceux affectés à chaque voix. Ce modèle ne prévoit donc pas de façon fiable le découpage morphologique de tous les mots, et il accepte des formes inconnues du malgache.

[2] L'actif-statif et le circonstanciel sont deux des valeurs que peut prendre la voix, une catégorie morphologique, au sens où on parle de voix active et passive en français. Lorsque le verbe passe de la voix active-stative à la voix circonstancielle, un complément circonstanciel passe parallèlement dans la position de sujet. Le malgache possède cinq voix (Ranaivoarison, 2016, p. 98).

[3] Et quelquefois un préfixe formatif de noms exprimant une habitude (Nhab) comme *mpidanadana* « ce qui reste habituellement ouvert ».

[4] Ces informations linguistiques ont été reprises telles quelles pour formaliser la catégorie grammaticale des noms. En effet, elles ont été suffisamment complètes, explicites et cohérentes pour pouvoir les utiliser dans le traitement automatique des langues.

Nous avons choisi de combler cette lacune en recensant systématiquement, d'une part, les variations morphologiques des lemmes, et d'autre part les combinaisons d'affixes avec ces variantes. Dans la pratique, notre description formelle prend la forme de deux activités : la codification de propriétés (catégorie grammaticale, combinaison d'affixes, variation de formes) et la construction de graphes (transducteurs de flexion et grammaires locales). Avant d'aborder la construction des graphes (section 4.), la codification effectuée pour construire le DE des noms est d'abord présentée dans cette section. Premièrement, la codification des catégories grammaticales et valeurs des préfixes formatifs de noms est abordée. Puis sont abordées respectivement la codification des combinaisons des affixes (classes affixales) et des variations de formes des radicaux (classes radicales).

3.1 Codification des catégories grammaticales et valeurs des préfixes formatifs de noms

Les catégories grammaticales et sémantiques qui entrent dans la construction du DE des noms issus de verbes sont listées ci-dessous.

PFN Préfixes formatifs de noms
PV Préfixes de voix
SV Suffixes de voix
V Verbes
:g noms d'agent et de profession
:m noms de manière et d'état
:n noms d'instrument
:t noms d'action

3.2 Codification des classes affixales

Une classe affixale est une classe de lemmes qui ont en commun la façon dont ils se combinent avec des affixes. Les codes de classes affixales des noms sont composés de trois cases.

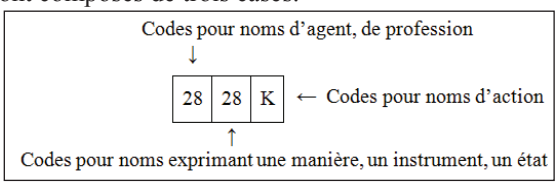


Figure A : Schéma général des codes de classes affixales des noms

- La première case indique les noms d'agent et de profession à préfixe *mp-* + préfixe de l'actif-statif, comme *mpanendy* analysé *mp-an-endy* « celui qui fait frire », *mpanjono* analysé *mp-an-jono* "pêcheur", *mpamboly* analysé en *mp-am-boly* « celui qui plante, jardinier, cultivateur ».
- La deuxième case indique les noms de manière, d'instrument et d'état à préfixe *f-* + préfixe de l'actif-statif, comme *fanadala* analysé *f-an-adala* "manière de duper", *fitaratra* analysé *f-i-taratra* "miroir", *fangatsiaka* analysé *f-an-gatsiaka* « l'état de celui qui a froid, de ce qui est froid »
- La troisième case est celle des noms d'action à préfixe *f-* + préfixe du circonstanciel, comme *fivahinianana* analysé *f-i-vahinian-ana* "action de voyager, de séjourner", *fanendasana* analysé *f-an-endas-ana* « action de faire frire, de rôtir, de griller ; poêle, marmite, rôtissoire », *fanabeazana* analysé en *f-ana-beaz-ana* « action d'agrandir, d'augmenter, d'élever, d'éduquer ».

La première et la deuxième cases ne peuvent recevoir que des chiffres et la troisième case des lettres en majuscules. Le code « v » est le seul utilisé pour chacune de ces trois cases si la case indique une absence de termes de noms d'agent, de profession, de manière, d'instrument, d'état ou d'action. Ci-dessous, ces types d'informations sont développés dans cet ordre.

3.2.1 Codes de noms d'agent et de profession à préfixe *mp-* + actif-statif

Les noms à préfixe *mp-* sont obtenus par alternance du préfixe de temps[5] avec *mp-,* préfixe formatif de nom d'agent et de profession. D'une manière générale, ces éléments obtenus par alternance reposent sur la voix active-stative comme dans *manjono* **V.act.-stat.** « pêcher » / *mpanjono* **N.** « pêcheur », *miady* **V.act.-stat.** « combattre » / *mpiady* **N.** « guerrier, combattant », *manafaingana* **V. act-stat.** « accélérer » / *mpanafaingana* **N.** « celui qui accélère ». Ci-après les codes de combinaison des affixes de l'actif-statif se combinant avec *mp-*.

Codes	Affixes
1	*Ø-*
2	*i-*
3	*an-*
4	*ana-*
7	*i-/an-*
21	*i-/an-/ana-*
23	*a-*
26	*i-/am-*
28	*i-/ana-*
30	*an-/ana-*
32	*ana-/anka-*
37	*anam*

Codes	Affixes
38	*i-/Ø-*
43	*am-*
60	*i-/am-/ana-*
61	*anka-*
62	*aha-*
63	*am-/ana-*
65	*an-/ana-/ian-*
66	*i-/a-*
67	*a-/anka-*
68	*Ø-/an-*
69	*i-/anam-*

Tableau 1 : Codes utilisés pour les noms formés sur l'actif-statif

Ces codes se placent en première position dans la chaîne des codes et sont composés uniquement de chiffre.

Si le radical à l'origine des noms ne fournit pas de noms d'agent et de profession alors un code "v" est utilisé pour marquer cet absence comme pour *mòndra* "épuiser une terre par une incessante production" qui

[5] Les préfixes de temps dont il s'agit ici sont ceux combinables avec l'actif-statif, c'est-à-dire /*m-* « présent » : *n-* « passé » : *h-* « futur »/ comme pour *lèha* « marcher » : *mandeha* au présent, *nandeha* au passé et *handeha* au futur.

a pour code v4E, la langue n'atteste pas l'existence du nom d'agent ou de profession **mpanamondra* mais fournit les formes comme *fanamondra* "manière d'épuiser la terre par une incessante production" (Nman) et *fanamondrana* "action d'épuiser la terre" (Nact) .

3.2.2 Codes de noms de manière, d'instrument, d'état à préfixe *f-* + actif-statif

Les mêmes codes de l'actif-statif utilisés au 3.2.1 sont utilisés pour former les noms de manière, les noms d'instrument et les noms exprimant un état. Les noms comme *fijery* **N.** « manière de regarder » issu de *mijery* **V. act.-stat.** « regarder », *fiendrinendrina* **N.** « l'état de stupidité » issu de *miendrinendrina* **V. act.-stat.** « être stupide », *fihogo* **N.** « peigne » issu de *mihogo* **V. act.-stat.** « peigner, se peigner » sont respectivement des noms exprimant une manière, un état, un instrument. En effet, les deux formations, l'une avec *mp-* et l'autre avec *f-* reposent toutes deux sur les affixes de l'actif-statif. Il s'ensuit que cette deuxième case est renseignée également par les chiffres présentés au tableau 1.

Si cette deuxième case n'est pas renseignée pour une entrée donnée alors elle est renseignée par le code "v" comme pour *hèry* 2 "1. A. Être fort, courageux, puissant, brave, zélé, faire bien, faire beaucoup. B. Gagner, l'emporter, vaincre, avoir un excédent, un surplus. 2. Rendre fort, fortifier, encourager. 3. Devenir fort, se fortifier, prendre courage" qui a pour code 67vXX, la langue n'atteste pas l'existence des noms de manière ou d'état **fahery* ou **fankahery* mais fournit les formes *mpahery* "habituellement vainqueur, un brave" (Nhab) [6] , *mpankahery* "celui qui fortifie" (Nag), *faherezana* "le courage, la force, la vigueur, l'entrain" et *fankaherezana* "action de fortifier" (Nact).

3.2.3 Codes de noms d'action à préfixe *f-* + circonstanciel

Les noms d'action sont formées sur le préfixe *f-*, préfixe formatif de noms, se combinant avec les affixes du circonstanciel comme *filalaovana* **N.** « action de jouer » issu du circonstanciel *ilalaovana* **V. circ.** « jouer », *fanadihadiana* **N.** « action de scruter, information » issu du circonstanciel *anadihadiana* **V. circ.** « scruter », *fieritreretana* **N.** « action de réfléchir » issu du circonstanciel *ieritreretana* **V. circ.** « réfléchir » . Ils sont obtenus par alternance du préfixe de temps[7]avec *f-*. Les codes des préfixes de la voix circonstancielle sont résumés dans le tableau ci-contre.

Codes	Affixes
A	*Ø-*
B	*i-*
C	*am-*
D	*an-*
E	*ana-*
F	*Ø-/an-*
G	*aha-*
H	*i-/Ø-*
I	*i-/am-*
J	*i-/an-*
K	*i-/ana-*

Codes	Affixes
L	*am-/ana-*
N	*an-/ana-*
O	*i-/an-/ana-*
S	*i-/am-/ana-*
T	*i-/an-/aha-*
U	*a-*
W	*i-/a-*
Z	*i-/anam-*
CC	*ana-/anka-*
XX	*a-/anka-*
ZZ	*an-/ana-/ian-*

Tableau 2 : Codes utilisés pour les noms formés sur le circonstanciel

Si cette troisième case n'est pas renseignée pour une entrée donnée alors elle est renseignée par le code "v" comme pour *zò* "tomber sur" qui a pour code 33v, la langue n'atteste pas l'existence du nom d'action **fanjoana* mais fournit les formes *mpanjo* "ce qui tombe sur" (Nag) et *fanjo* "manière de tomber sur" (Nman).

3.3 Codification des classes radicales

Une classe radicale est une classe de lemmes qui ont en commun la façon dont varie leur radical. Les codes de classes radicales des noms sont composés de trois cases comme pour les verbes (Ranaivoarison, 2016, p. 218). Ces mêmes codes de classes radicales employés pour les verbes sont réutilisés car les noms sont également issus de verbes. Ci-dessous les principes utilisés pour ces codes sont résumés.

- La première case désigne les finales des radicaux verbaux qui peuvent être « 0 », « 1 », « 2 » ou « 3 ».
- La deuxième case désigne la compatibilité des radicaux verbaux avec le suffixe *-ina* et peuvent être « a » ou « i ».
- La troisième case indique les phénomènes[8] qui peuvent apparaître au niveau des radicaux verbaux lorsque ceux-ci sont entrent en contact avec les affixes.

Les codes des classes radicales sont introduites par la lettre V désignant les verbes. Ils sont aux alentours de 170 correspondant à des transducteurs de flexion (4.1) qui permettent de générer les paradigmes flexionnels et les relier aux affixes.

4 Construction des graphes de noms

Deux types de graphes sont associés aux codes de classes affixales et codes de classes radicales. Ces deux types de graphes sont présentés ci-après en exa-

[6] Voir note 3.

[7] Les préfixes de temps dont il s'agit ici sont ceux combinables avec le circonstanciel, c'est-à-dire *Ø-* « présent »/*n-* « passé »/*h-* « futur » comme pour *lèha* « marcher » : *andehanana* au présent, *nandehanana* au passé et *handehanana* au futur.

[8] Ces phénomènes sont par exemple de phénomènes de suppression ou de remplacement de la première lettre d'un radical, d'insertion d'une lettre au début ou d'utilisation d'un élargissement, etc.

minant premièrement ceux qui sont rattachés aux codes de classes radicales et deuxièmement ceux rattachés aux codes de classes affixales.

4.1 Transducteurs de flexion

Les transducteurs de flexion sont les graphes qui se rattachent aux codes de classes radicales. Ils fournissent à l'aide du programme de génération de formes d'Unitex les variantes morphologiques des radicaux formant des noms. Pour un radical comme *lèha* « marcher » par exemple, le transducteur de flexion V0ibe permet de générer automatiquement les variantes morphologiques de *lèha* comme *dèha*[9] dans *mpandeha* « voyageur, passant » ou dans *fandeha* « manière de marcher, démarche », et comme *dehán* dans *fandehanana* « action de marcher, marche, chemin » en indiquant les affixes qui vont avec les variantes. Ci-après, le graphe de transducteur de flexion V0ibe est fourni.

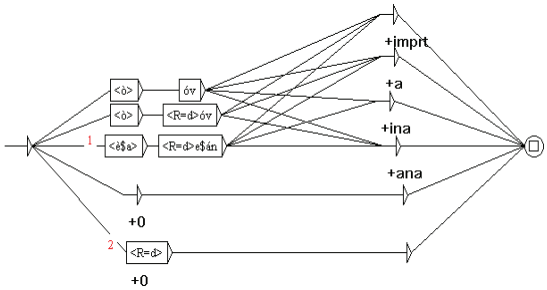


Figure B : Transducteur de flexion V0ibe

Les chemins 1 et 2 dans la figure B. permettent de générer les formes *dehán* et *dèha* ; ils indiquent respectivement les affixes avec lesquels ils se combinent. Le chemin 1 fournit par exemple la forme *dehán* et lui associe une propriété codée **+ana** indiquant qu'il se combine avec l'affixe *-ana* et se retrouve dans la forme *fandehanana* « action de marcher, marche, chemin » pour les noms. Les autres propriétés (**+imprt**, **+a**, **+ina**) pour ce chemin sont utilisées pour les formes verbales (Ranaivoarison, 2016, p. 227.). La boîte avec **+0** indique qu'après la variante morphologique il n'y a plus de suffixe comme dans le chemin 2 (Fig. B). En effet, après la variante morphologique *dèha*, il n'y a plus de suffixe, comme dans les formes nominales *mpandeha* « voyageur, passant » et *fandeha* « manière de marcher, démarche ».

4.2 Graphes de grammaires locales

Dans l'état actuel de notre recherche, 67 graphes de grammaires locales ont été créés. Ils correspondent aux codes de classes affixales (3.2). Ces graphes permettent l'analyse morphologique des noms issus des verbes. Ci-contre, le graphe de grammaire locale v2B pour les radicaux verbaux qui n'ont pas de noms d'agent ni de profession mais ont toutes les autres formes nominales (noms de manière ou d'état et noms d'action) est fourni.

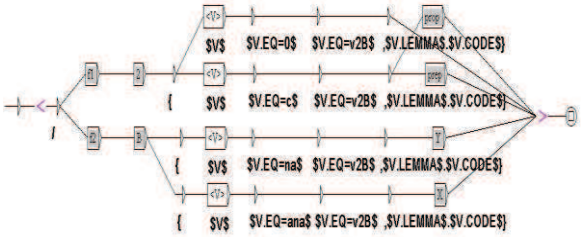


Figure C : Graphe de grammaire locale v2B

Comme exemple se rattachant à ce graphe, nous avons *zozozòzo* « bourdonner, bruire, siffler », d'où *fizozozozo* « manière de bourdonner, de bruire, de siffler » (Nman), *fizozozozoana* « bourdonnement, bruissement, sifflement » (Nact). Ce type de graphe peut aussi être utilisé par des programmes de génération de formes non plus pour découper les formes reconnues mais pour construire, indépendamment d'un corpus donné, des listes de formes nominales. Rakotoalimanana (2000, p. 378) expose un exemple de ce programme de génération de formes avec les formes verbales. Il y présente un prototype d'Analyseur – Générateur des Termes prédicatifs Malgaches (AGTM) implémenté en langage Prolog.

5 Les dictionnaires de noms

Les codes de classes affixales et radicales sont insérés dans le dictionnaire de noms et opèrent directement sur le dictionnaire à l'aide des transducteurs de flexion et des graphes de grammaire locale. Dans cette section, le dictionnaire électronique des noms issus de verbes (DEMA-NVS) est présenté en premier lieu ; ensuite, le dictionnaire des variantes morphologiques des radicaux (DEMA-NVSflx) est abordé en second lieu.

5.1 DE des noms issus des verbes (DEMA-NVS)

Les entrées du DEMA-NVS sont les radicaux verbaux. Dans l'état actuel de notre recherche, elles sont au nombre de 1500 ; toutes les entrées commençant par A – J, M, N, Z ont été codées. Ci-après un extrait de ce dictionnaire.

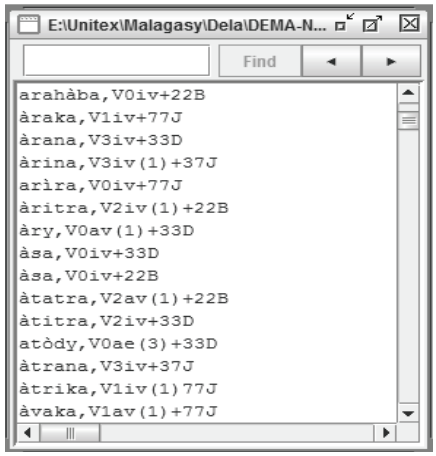


Figure D : DEMA-NVS

[9] L'accent graphique note une information fournie par le dictionnaire sur l'accent tonique, mais il n'est généralement pas précisé dans les textes écrits.

Avec les conventions d'Unitex, les articles du dictionnaire sont séparés des entrées par une virgule et les codes après la virgule sont les articles du dictionnaire. Les avantages d'un dictionnaire construit par les linguistes sont qu'il est précis et facile à mettre à jour. Il fournit des informations jugées pertinentes soit pour les futurs programmes d'aide à la construction de dictionnaires usuels, soit pour les logiciels et applications destinées aux utilisateurs finaux.

5.2 DE des variantes morphologiques des noms (DEMA-NVSflx)

Les variantes morphologiques des radicaux verbaux formant des noms sont rangées dans un autre dictionnaire appelé DEMA-NVSflx. À proprement parler, le dictionnaire n'est pas un dictionnaire de formes fléchies de noms, il est un dictionnaire morphologique contenant les variantes morphologiques du radical, et indiquant par des codes les affixes se combinant avec ces variantes. Ci-après une image du DEMA-NVSflx.

Figure E : DEMA-NVSflx

Pour fournir un véritable dictionnaire de formes fléchies, un autre programme de génération automatique de termes est requis. Il servira plus tard à de nombreuses applications pratiques.

6 Test du dictionnaire

Des procédures d'évaluation du dictionnaire ont été mises au point sur un extrait du corpus journalistique du malgache contemporain (cjmc) de Diwersy (2009) qui n'a pas été utilisé pour construire le dictionnaire. Le dictionnaire a été testé sur les 50 premières phrases du cjmc1[10] qui comportent 35 noms différents. Parmi les 35 noms différents :

- 6 sont reconnus et découpés par Unitex en utilisant les ressources codées
- 29 ne sont pas reconnus car ils sont absents du dictionnaire. Parmi ces derniers :
    - Toutes les classes radicales existent déjà dans les ressources
    - Pour les classes affixales :
        - 24 noms non-reconnus correspondent en réalité à 6 classes affixales existantes dans le fichier des grammaires locales
        - pour les 5 autres noms non-reconnus, les classes affixales sont à insérer dans les ressources

En termes de classes radicales, le texte est à 100% couvert tandis qu'en termes de classes affixales, il est à 86% couvert. D'une manière générale, la plupart des classes radicales et affixales des radicaux ont déjà été construites dans Unitex au cours du travail. Il s'agit ensuite d'enrichir le dictionnaire de radicaux verbaux et le dictionnaire peut couvrir le lexique des noms issus de verbes.

7 Couverture lexicale

Une fois que le dictionnaire est enrichi des radicaux verbaux formant des noms, Unitex est capable de faire les analyses morphologiques des noms d'agent, de profession, de manière, d'instrument, d'état et d'action dérivés de ces radicaux. Il peut reconnaître également d'une part ces noms couplés avec des pronoms personnels du type *fijeriko* « mon regard », *filalaoko* « ma manière de jouer », *fisaorako* « mon remerciement » ou avec des prépositions comme *mpamilin'* « le chauffeur de » et d'autre part les variantes morphologiques de ces noms au début des radicaux au contact d'un trait d'union comme *pifamoivoizana* (de *fifamoivoizana* « action de circuler, circulation ») dans *lozam-pifamoivoizana* « accident de la circulation » dans les mots composés. Les transducteurs de flexion et les graphes de grammaires locales construits fonctionnent correctement et le codage des entrées pour constituer un DE complet des noms issus des verbes est en cours. Si dans l'état actuel de notre recherche, nous sommes à 1500 entrées de ce dictionnaire, il reste 53% des entrées qui ont besoin d'être insérées dans le dictionnaire. Une fois l'enrichissement du dictionnaire complet, un dictionnaire DEMA-NVS des noms issus de verbes du malgache sera disponible, ce qui augmentera d'une manière assez considérable la couverture lexicale du malgache.

8 Conclusion

La construction de dictionnaire électronique des noms issus de verbes est en phase de constitution au Centre Interdisciplinaire de Recherche Appliquée au Malgache. S'il reste des entrées manquantes qui doi-

[10] Cjmc 1 est une partie du corpus journalistique du malgache contemporain de Diwersy (2009) dont nous avons divisé en quatre parties (voir Ranaivoarison, 2016, p. 260). Cjmc1 comporte 180 000 mots et 12 700 phrases.

vent être insérées dans le dictionnaire pour constituer un dictionnaire complet, ce dictionnaire est déjà utilisable pour certaines applications. Une fois que la construction de ce dictionnaire sera terminée, la construction des dictionnaires de noms issus d'adjectifs et de noms simples constituerait les prochaines priorités pour former un dictionnaire de noms simples qui tend à l'exhaustivité du vocabulaire.

L'extension de ce dictionnaire aux autres catégories grammaticales (adjectifs, adverbes, et les autres catégories à faible variation de formes telles que les conjonctions, les prépositions, etc.) permettra d'avoir un dictionnaire morphologique électronique complet du malgache qui servira d'accès aux dictionnaires de mots composés et d'un lexique-grammaire représentant systématiquement les propriétés syntaxiques des mots de la langue. Ces informations seront ensuite utilisées dans d'autres programmes informatiques qui ont pour finalité la génération de formes, la normalisation, la correction orthographique et/ou grammaticale. En d'autres termes, elles serviront à la construction d'outils de TAL performants et accessibles aux grands publics.

Références

Berment, V. (2004). *Méthodes pour informatiser des langues et des groupes de langues « peu dotées »*. Thèse de doctorat. Université Jean Fourier, Grenoble 1.

Dalrymple, M., Liakata, M., Mackie, L. (2006). Tokenization and morphological analysis for Malagasy. In: *Computational Linguistics and Chinese Language Processing* 11 (4), pp. 315-332. Taipei: Institute of Linguistics, Academia Sinica.

Diwersy, S. (2009). *Corpus journalistique du malgache contemporain.* Romance Philology Department University of Cologne.

Gross, M. (1989). La construction de dictionnaires électroniques. In : *Annales des télécommunication, tome 44* N°1, 2. Issy-les-Moulineaux/lannion : CNET.

Koskenniemi, K. (1983). *Two-Level Morphology: A general Computational Model for Word-Form Recognition and Production.* Department of General Linguistics, University of Helsinki.

Koskenniemi, K. and Church, K.W. (1988). Complexity, two-level morphology and Finnish. In: *COLLING'88.*

Nam, J. S. (1994). Construction d'un lexique électronique des noms simples en coréen. In : *Lexiques-grammaires comparés et traitements automatiques*. Université du Québec à Montréal : Jacques Labelle, pp. 219-245.

Nam, J. S., Paumier, S. (2014). Un système de dictionnaire de mots simples du coréen. Fryni Kakoyianni-Doa. Penser le Lexique-Grammaire. Perspectives actuelles, Honoré Champion, pp.481-490, 2014, Collection Colloques, congrès et conférences. Sciences du Langage, histoire de la langue et des dictionnaires. 30th International Conference on Lexis and Grammar (Nicosia, Cyprus, 2011), 978-2-7453-2512-9.

Oflazer, K. (1993). Two-level Description of Turkish Morphology. In: *EACL'06.* Netherlands, Utrecht.

Paumier, S. (2016). *Unitex 3.1. Manuel d'utilisation.* Université Paris-Est Marne-la-Vallée. Version française.

Rajaona, S. R. (1972). *Structure du malgache.* Antananarivo : Ambozontany.

Rakotoalimanana, H. D. (2000). *Structure morpho-syntaxique et modélisation informatique.* Thèse de doctorat. Université Nancy 2.

Ranaivoarison, J., Laporte, É., Ralalaoherivony, B. S. (2013). Formalisation of Malagasy conjugation. In: *Language and Technology Conference*. Poznan, Poland. pp.457-462.

Ranaivoarison, J. (2015a). *Description du dictionnaire électronique des verbes simples du malgache*. Session Poster. Colloques Jeunes Chercheurs. Montpellier.

Ranaivoarison, J. (2016). *Construction de dictionnaire électronique des verbes du malgache*. Deutschland : Editions Universitaires Européennes.

Annotation d'éléments spatialisés dans l'oral transcrit

Hélène Flamein
Laboratoire Ligérien de Linguistique (LLL, UMR 7072)
Université d'Orléans
helene.flamein@univ-orleans.fr

Résumé

Dans le domaine du Traitement Automatique des Langues (TAL), les travaux sur des données spatialisées sont de plus en plus nombreux et présentent de nouveaux enjeux. Cette communication propose une réflexion sur les caractéristiques propres à la dénomination d'un lieu dans le corpus ESLO (Enquête SocioLinguistique à Orléans). Les noms de lieux sont soumis à variation d'un locuteur à l'autre. Avant de proposer une annotation automatique des lieux qui prendrait en compte ces variations, il est nécessaire de s'interroger sur la typologie des balises à utiliser. Les conventions d'annotation établies aideront à la constitution d'un corpus de référence, composant indispensable dans l'élaboration ou l'évaluation d'un système d'annotation automatisé.

Mots clés :

Désignation de lieux, Lieux subjectifs, Conventions d'annotation, Traitement Automatique du Langage, Entités nommées, ESLO, Corpus oral

1 Introduction

Cet article s'inscrit dans le cadre d'un travail de thèse portant sur l'expression de la subjectivité dans l'oral spontané. L'objectif général de ce travail est de permettre l'analyse automatique de la perception de la ville d'Orléans par ses habitants grâce à l'exploitation du corpus ESLO2. Cette analyse est fondée sur une succession d'annotations et la première d'entre-elles concernera l'identification de toutes les mentions de lieux présentes dans le corpus. Afin de traiter l'ensemble des données disponibles, nous utilisons les techniques du Traitement Automatique des Langues (TAL). Les lieux et les expressions subjectives relatives à ces lieux sont détectés automatiquement et analysés par la suite pour observer la variation de la perception des lieux par les différents locuteurs. Enfin, les résultats de cette analyse prendront la forme d'une carte représentant les lieux identifiés avec les déclarations des locuteurs interrogés relatives à ces lieux pour présenter le portrait de la ville d'Orléans.

La recherche présentée ici se concentrera sur le travail préparatoire à la détection automatique des mentions de lieux et plus particulièrement à la tâche d'annotation manuelle de ces entités. Les spécificités du corpus et des données à identifier seront dans un premier temps présentées. La méthodologie de la constitution du corpus de référence sera ensuite explicitée avec une attention particulière aux conventions d'annotations utilisées.

2 Présentation des données

2.1 Le corpus ESLO

Cette étude est fondée sur le corpus ESLO[1] (Enquête SocioLinguistique à Orléans) (Eshkol-Taravella et al. 2012), projet du Laboratoire Ligérien de Linguistique, qui met au cœur de son investigation les pratiques langagières dans la ville d'Orléans. Il se décompose en deux séries d'enquêtes, ESLO1 et ESLO2 qui cumulent ensemble près de 700h d'enregistrements. La première campagne ESLO1 initiée par des linguistes anglais avait pour objectif de présenter le français tel qu'il était parlé. ESLO2 propose le même travail à 40 ans d'intervalle « en prenant en compte l'expérience d'ESLO1 et l'évolution des cadres théoriques et méthodologiques de la constitution et de l'exploitation de grands corpus

[1] http://eslo.huma-num.fr/

oraux à visée variationniste » (Baude et Dugua, 2011). Ces deux corpus comprennent différentes situations d'enregistrements : entretiens face à face, interviews de personnalités, enregistrements dans des cours de récréations, pendant des repas, etc.

La transcription des enregistrements suit un protocole très précis et détaillé dans le *Guide du Transcripteur et du Relecteurs des ESLOs* [2]. Chaque enregistrement est transcrit orthographiquement avec une distinction entre les tours de parole. La convention de transcription préconise de transcrire sans signes de ponctuation et sans majuscules. Les points d'interrogation pour les questions et les majuscules des noms propres sont les seules exceptions admises.

2.2 Modules sélectionnés pour l'analyse

En considérant le contexte d'énonciation des enregistrements et les trames qui ont servi à guider les entretiens, deux modules du corpus ESLO2 ont été sélectionnés : Entretiens et Itinéraires.

Les Entretiens consistent en une discussion en face à face entre un chercheur et un locuteur témoin. Le chercheur mène la discussion selon une trame préétablie qui reste assez souple pour laisser place à la spontanéité du discours du locuteur. D'une manière générale, la trame invite ce dernier à faire état de son histoire personnelle, à partager ses habitudes de vie, etc. Chacune des personnes enregistrées est un habitant d'Orléans ou de son agglomération.

Au total, le module Entretiens d'ESLO2 comprend 84 transcriptions pour un total de 150h et environ 1 166 660 mots.

Le module Itinéraires regroupe des enregistrements réalisés en pleine rue. Des étudiants ou chercheurs vont à la rencontre de piétons pour leur demander leur chemin jusqu'à la mairie comme dans l'exemple [1] ou jusqu'à un autre endroit connu d'Orléans.

1. FD720: *bonjour excusez-moi de vous déranger je cherche la mairie d'Orléans*
   MH315: *c'est vers la cathédrale à pied ?*
   FD720: *oui ou en tram ou en ce que vous voulez [rire] du moment que j'y arrive [rire]*
   (ESLO_iti_06_11_C)

La question est dans un premier temps posée à micro discret. Une fois que le locuteur a répondu, on lui révèle le micro et lui demande de reformuler sa réponse. Suivent quelques questions sur les habitudes du locuteur dans la ville et son avis sur celle-ci. La collecte a été effectuée dans divers endroits de la ville afin d'interroger des locuteurs représentatifs de la diversité sociologique de la ville. De par leur constitution, ces courts enregistrements forment un matériel riche en mentions de lieux relatives à la ville d'Orléans.

Au total, le module Itinéraires d'ESLO2 comprend 91 transcriptions qui représentent 5h d'enregistrements et environ 69 330 mots.

3 Détection automatique des lieux dans l'oral spontané

3.1 Etat de l'art

Selon (Fort, 2012) l'annotation en tant que pratique qui a cours en TAL « consiste à apposer des étiquette (ou notes) de nature linguistique ou reflétant l'usage des technologies du TAL sur du discours oral ou écrit ». Les annotations permettent un accès direct au contenu du corpus annotés et constituent la base des tâches d'extraction d'informations en TAL.

Depuis les années 1990 et la dernière série des conférences américaines MUC (Message Understanding Conferences), la question de la reconnaissance des entités nommées est incontournable dans le domaine du TAL. Selon (Ehrmann, 2008), les entités nommées représentent « toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Ces entités représentent des objets textuels porteurs de sens généralement classés selon plusieurs catégories : lieux, personnes, organisations, dates, unités monétaires et pourcentages (Chinchor, 1998 ; Maurel et al., 2011 ; Nadeau et Sekine, 2009). Dans le domaine du TAL, les noms de lieux sont considérés comme des entités nommées.

La tâche de reconnaissance des entités nommées (REN) est devenue une tâche indépendante qui est désormais au centre de différentes campagnes d'évaluations d'outils dédiés à l'extraction d'informations. Plusieurs campagnes comme ESTE2R[3] ou ETAPE[4] évaluent justement l'annotation des entités nommées dans des corpus d'émissions radiophoniques ou télévisuelles. En amont de l'évaluation de ces campagnes, des échantillons de corpus ont été annotés manuel-

[2] http://eslo.huma-num.fr/index.php/pagemethodologie?id=71

[3] http://www.afcp-parole.org/camp_eval_systemes_transcription/

[4] http://www.afcp-parole.org/etape.html

lement selon des conventions préétablies. Ces conventions présentent notamment des recommandations pour l'annotation des lieux.

Selon la définition du TLF[5], un *lieu* est une « portion déterminée de l'espace ». Cette portion peut être localisée, identifiée sur une carte au moyen de coordonnées géographiques. Lesbeguerrie (2007) présente l'idée d'entité spatiale qu'il précise selon deux catégories. Les entités spatiales absolue correspondent aux informations spatiales les plus « primitives » et les plus proches de la définition des entités nommées de type lieu (ex : la ville d'Orléans, Olivet). Les entités spatiales relatives allient entités nommées et indications spatiales. Des informations objectives comme le nom d'un lieu sont directement associées à des renseignements subjectifs à propos du lieu en question (ex : près de la ville d'Orléans, du côté d'Olivet). La subjectivité a déjà été liée à des notions géographiques. C'est le cas du projet Senterritoire[6] qui a comme objectif de détecter les opinions et les sentiments liés à l'aménagement d'un territoire

3.2 Interférences de la subjectivité dans la dénomination d'un lieu

Selon (Dominguès et Eshkol, 2013), l'écriture des toponymes « *fait appel à des règles complexes qui s'appuient sur des connaissances linguistiques et extralinguistiques* ». Dans des contextes d'écriture moins normalisés comme sur le Web (blogs, commentaires, réseaux sociaux), l'écriture des noms de lieux est rapidement soumise à de réelles variations propres à l'utilisateur observé (troncation, abréviation, faute de frappe, etc.). Ces variations sont souvent induites par la tradition orale que l'on retrouve donc dans le corpus étudié :

2. *« ah ben si tu peux redescendre tu prends la tu prends la rue qui est là et tu vas tout au bout jusqu'à la* ***rue de la Rép-*** *tu vois où elle est ? la* ***rue de la République*** *? »* (ESLO2_iti_06_11_C)

Dans cet exemple, le nom de la *rue de la République* est mentionné dans sa forme complète et dans une forme tronquée : *rue de la Rép-*.

3. « *parce que mon grand-père euh donc était boulanger ét- avait une boutique à* ***La Ferté-Saint-Aubin*** » (ESLO2_ENT_1025_C)

4. « *je passais pas* ***La Ferté*** *ça faisait loin hein ça me faisait cinquante kilomètres* » (ESLO2_ENT_1023_C)

Dans l'exemple [3] est mentionné La Ferté qui est la forme étendue du nom de la ville La Ferté-Saint-Aubin, mentionnée dans l'exemple [4].

La mention de lieux à l'oral présente des variations différentes dans lesquelles la perception du lieu peut transparaitre. Cette dénomination est un processus social réapproprié subjectivement et est déterminée par la personnalité, l'histoire, du locuteur. Un lieu peut être approprié ou apprécié, ou non, par un locuteur. Eshkol-Taravella et Flamein (à paraître) distingue deux types de perception des lieux dans l'oral spontané : la perception exprimée à travers la variation dans la dénomination d'un lieu par des locuteurs et la perception manifestée dans le contexte d'emploi des lieux.

5. *en gros euh sous* ***les Arcades*** (ESLO2_ENTJEUN_04_C)

Dans cet exemple [5], les Arcades sont le surnom donné à la *rue Royale*, une rue centrale à Orléans. Celle-ci est bordée sur toute sa longueur par des galeries à arcades. Cette spécificité architecturale a conduit les Orléanais à se référer à cette rue en substituant son nom officiel par une appellation plus imagée. On observe ainsi une véritable réappropriation du nom d'un lieu. Faire allusion à une entité en utilisant un surnom est un cas de personnalisation, d'appropriation d'un lieu par un locuteur. Dans l'exemple :

6. *b- c'est la grande région euh c'est* ***la grande région*** *euh* ***Centre*** (ESLO2_ENT_1034_C)

le locuteur emploie l'adjectif grande à propose de la région Centre. Le lexique évaluatif dans le contexte proche du nom du lieu constitue un indice sur la vision du lieu du locuteur.

L'enjeu de notre travail est de détecter toutes les mentions de lieux présentes dans le corpus tout en prenant en compte leur capacité à varier en fonction du locuteur afin de construire le portrait de la ville d'Orléans. Plus que de pouvoir détecter les formes tronquées ou abrégées d'une entité nommée, l'intérêt se trouve aussi dans la

[5] http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?12;s=668480715;r=1;nat=;sol=1;

[6] http://www.msh-m.fr/la-recherche/programmes-actuels/senterritoire/

possibilité de faire le lien entre la forme modifiée du nom du lieu et sa forme originelle. Le système responsable de l'annotation automatique doit être capable de faire le lien entre une entité nommée, nom officiel du lieu, et ses possibles variantes grâce à l'observation du niveau d'analyse de la perception intrinsèque au nom du lieu observé. Ce lien permettra aussi de rendre géolocalisable sur une carte chacun des lieux identifiés, qu'ils soient mentionnés via leurs noms officiels ou via une variante de celui-ci.

Nous allons donc présenter la méthodologie employée en préparation de l'automatisation de l'annotation des mentions de lieux dans l'oral spontané.

4 Protocole d'identification des mentions de lieux

4.1 Constitution d'un corpus de référence

Que ce soit dans l'optique de la création d'un nouveau système d'extraction d'information ou pour toute utilisation d'un système existant, il est nécessaire d'évaluer les performances de ce système. Les mesures de Rappel, Précision et F-mesure assurent l'évaluation des performances du modèle choisi. Ces mesures s'appuient sur la comparaison d'un corpus annoté automatiquement par le modèle à évaluer et un corpus de référence. Ce corpus de référence doit correspondre à un échantillon du corpus général dans lequel toutes les données à identifier sont toutes annotées manuellement et prêtes à être extraites.

En l'occurrence, nous avons sélectionné 5 transcriptions dans les modules Entretiens et Itinéraires afin de constituer notre propre corpus de référence.

Transcriptions	Durée	Nombre de mots
ESLO_ENT_1059	1:40:00	19 449
ESLO_ENT_1002	1:37:00	14 791
ESLO_ENT_1034	1:30:00	15 788
ESLO_iti_08_04	0:06:40	1001
ESLO_iti_02_09	0:04:00	299
Totaux	4:57:40	51328

Tableau 1 : Volume de données par transcriptions

Cet échantillon annoté manuellement en lieux sera la référence pour l'évaluation de notre système. L'annotation est fondée sur des conventions d'annotations établies en fonction des besoins propres à notre analyse.

4.2 Conventions d'annotation

Comme abordé précédemment, les entités nommées classiques (cf. [3.1]) et celles soumises à variations (cf. [3.2]) seront considérées dans l'annotation. Celle-ci se fera au moyen de la balise XML **<loc>** … **</loc>** et devra comprendre les informations suivantes en attributs de la balise principale :

4.2.1 Le type de lieu

La typologie des lieux participera à une première catégorisation des mentions identifiées. Cette information permettra de traiter différemment certaines annotations au moment de l'analyse de la subjectivité : le nom d'une ville sera traité différemment de celui d'une rue ou d'une structure à but éducatif par exemple.

Les conventions d'annotation des entités nommées de type lieu présentées ici s'inspirent notamment de celles établies pour la campagne ETAPE [7] (Rosset, Grouin et Zweigenbaum, 2011). Ce projet avait pour objectif d'évaluer les performances des technologies vocales appliquées à l'analyse de flux télévisés en langue française. Les conventions d'annotations des entités nommées Quaero utilisées dans ce projet propose de classer les lieux selon la typologie suivante :

Lieux administratifs	
Ville/quartier	loc.adm.town
Région	loc.adm.reg
Pays	loc.adm.nat
Supranational	loc.adm.sup
Lieux physiques	
Terrestres	loc.phys.geo
Aquatiques	loc.phys.hydro
Astronomiques	loc.phys.astro
Voies	
Voies	loc.oro
Bâtiments	
Bâtiments	loc.fac
Adresses	
Adresses postales	loc.add.phys
Adresses elec/tel/fax	loc.add.elec

Tableau 2 : Typologie des entités nommées de type lieu selon Quaero

Les entités nommées de type lieu sont très proche de celle considérées comme des organisations. Ici, nous considérons que, pour une entité nommée normalement catégorisée comme une organisation, l'information de la localisation pré-

[7] http://www.afcp-parole.org/etape.html

vaut sur celle de la fonction de l'entité. Ainsi, toutes les organisations seront annotées comme des lieux dans notre corpus. Pour ce faire, nous nous référons aussi aux conventions d'annotation décrites lors la campagne ESTER2[8], projet antérieur à ETAPE avec des objectifs similaires de mesure de performances de systèmes de transcriptions d'émissions radiophoniques. Dans ces conventions, les organisations sont réparties dans les catégories suivantes :

Organisations	
Politique	org.pol
Educative	org.edu
Commerciale	org.com
Non commerciale	org.non-profit
Média & divertissement	org.div
Géo-socio-administrative	org.gsp

Tableau 3 : Typologie des entités nommées de type organisation selon ESTER2

A partir de ces deux conventions, nous avons proposons de typer les entités nommées identifiées de la façon suivante :

<loc type=" ">	
type ="ville"	Villes
Orléans, Paris, La Ferté-St-Aubain…	
type ="pays"	Pays
France, Espagne, Royaume-Uni, Chine…	
type ="voie"	Rues, avenues, ponts…
rue de la République, Pont Royal…	
type ="naturel"	Lieux physiques naturels
Forêt d'Orléans, Loire,…	
type ="monument"	Lieux à dimension historique, touristique
Cathédrale Sainte Croix, Hôtel Groslot…	
type ="admin"	Fonction administrative
Mairie d'Orléans, Office du Tourisme, CAF…	
type ="educatif"	Fonction éducative
Lycée Pothier, Université d'Orléans…	
type ="commerce"	Fonction commerciale
Carrefour, H&M, Menphis Coffee…	
type ="ncommerce"	Fonction non commerciale
Hôpital de la Source, Secours Populaire,…	

Tableau 4 : Nouvelle typologie des lieux

Cette typologie conserve les catégories principales proposées par Quaero en ce qui concerne les lieux que l'on peut découper administrativement (comme les villes, pays, etc.). Par rapport aux conventions d'ESTER2, les lieux avec une fonction d'organisation sont typés de façon similaire. Toutefois, selon les conventions d'ESTER2, le type « politique » représente les organisations à caractères politiques telles que les organisations qui s'occupent des affaires gouvernementales (partis politiques, mairies, minsitères, etc.) ou les organisations militaires reliées au gouvernement (ex : CIA, Marine Nationale…), etc. Nous ne conservons pas ce type puisque nous considérons que les entités comme les partis politiques ou organisations militaires ne sont pas assimilables à des lieux. Si des lieux à fonction politique sont évoqués, ils seront plutôt inclus avec le type « admin » de notre convention.

4.2.2 Zone géographique :

Trois zones géographiques sont distinguées dans l'annotation. Celles-ci différencient les lieux situés à Orléans, les lieux hors Orléans mais situés dans son agglomération et les lieux en dehors de l'agglomération (cf. Tableau 5). Le découpage de ces trois zones correspond aux découpages administratifs de la ville d'Orléans et de son agglomération.

<loc type=" " zone=" ">	
zone ="0"	lieux hors agglomération orléanaise
Paris, Tours, Indre, Bretagne, Rhône, Seine …	
zone ="1"	lieux hors Orléans mais inclus dans l'agglomération
Saint Jean de la Ruelle, Saran, Auchan…	
zone ="2"	lieux situé à Orléans
Orléans, rue de Bourgogne, Key-West…	

Tableau 5 : Zone géographique

[8] http://www.afcp-parole.org/camp_eval_systemes_transcription/

L'information de la zone géographique permet des traitements différents entre les annotations. Par exemple, un lieu considéré hors agglomération orléanaise n'aura pas à être géoréférencé sur la carte finale.

7. *« c'est pas ça pose pas de problème donc euh ce qui manque à* **<loc type="ville" zone="2" label="Orléans">*Orléans*</loc>** *je dirais tu peux l'avoir à* **<loc type="ville" zone="0" zone="Paris">*Paris*</loc>** *donc c'est vrai que euh »* (ESLO2_ENT_1008_C)

Si un lieu est identifié comme appartenant à la zone d'Orléans comme dans l'exemple [7], alors on interrogera son contexte proche pour analyser les éventuelles marques de perception présentes. Si un autre lieu est présent dans ce contexte, il ne sera pas considéré de la même façon s'il fait partie ou non de la même zone.

4.2.3 Nom officiel du lieu

L'attribut label trouve son intérêt dans la tâche de géolocalisation du lieu identifié. La valeur de l'attribut sera le nom officiel du lieu identifié. Cette information servira à rechercher dans une base de données les coordonnées GPS du lieu pour le placer sur la carte finale. Dans ces exemples mentionnés précédemment, on annotera :

8. *« ah ben si tu peux redescendre tu prends la tu prends la rue qui est là et tu vas tout au bout jusqu'à la* **<loc type="voie" zone="2" label="rue de la République">*rue de la Rép-*</loc>** *tu vois où elle est ? la* **<loc type="voie" label=" rue de la République">*rue de la République*</loc>** *? »* (ESLO2_iti_06_11_C)

9. *« je passais pas* **<loc type="ville" zone="0" label="La Ferté-Saint-Aubin">*La Ferté*</loc>** *ça faisait loin hein ça me faisait cinquante kilomètres »* (ESLO2_ENT_1023_C)

Le nom officiel d'un lieu correspond à sa forme complète, sans aucune modification. Un moyen de vérifier cette donnée est de se référer à des dictionnaires ou à des bases de données spécialisées dans les noms de commerces ou des bases Linked Open Data comme Geonames pour les noms de villes, de pays, etc.

5 Conclusion et perspectives

Les entretiens enregistrés portent sur la ville d'Orléans. Les locuteurs parlent de leur ville et, par conséquent, mentionnent dans leurs discours les différents lieux. La nature orale du corpus et la diversité des locuteurs favorisent les variations dans les désignations de lieux. Des variations à prendre en compte afin de permettre la détection automatique de ces entités.

Le repérage et l'annotation des lieux entre dans une démarche globale visant l'étude de la perception des lieux par les habitants d'Orléans. Nous proposons une procédure d'annotation manuelle de ces lieux dans une transcription tout en tenant compte des multiples variations dans leur désignation. Cette ressource de référence prépare l'élaboration et l'évaluation du système d'annotation automatique des lieux et de leur perception dans l'oral spontané transcrit.

A terme, les résultats de l'analyse complète seront représentés cartographiquement. D'une part, les énoncés des locuteurs seront géoréférencés en fonction du lieu mentionné. D'autre part, les données issues des bases de données Linked Open Data comme Wikipedia seront associées aux énoncés pour contraster la perception exprimée avec une image objective du lieu à décrire.

L'association des témoignages et des données objectives donnera à cette carte une dimension anthropologique, sociologique et offrira la possibilité de constituer un véritable portrait sonore d'Orléans.

6 Références

BAUDE O., DUGUA C. (2011) *(Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?*. Corpus, 2011, Varia, 10, pp.99-118.

CHINCHOR N. (1998). *Overview of MUC-7*. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998.

DOMINGUES C., ESHKOL-TARAVELLA I. (2015). *Toponym recognition in custom-made map titles*. International Journal of Cartography, Volume 1, Taylor & Francis.

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthode de désambiguïsation.* PhD thesis, Université Paris 7.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER, I. (2012). *Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012*. in Ressources linguistiques libres, TAL. (vol. 52, n° 3, p. 17-46).

FORT Karën (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Traitement du texte et du document. Université Paris-Nord - Paris XIII, 2012. Français.

FORT K., EHRMANN M., NAZARENKO A. (2009). *Vers une méthodologie d'annotation des entités nommées en corpus ?* Traitement Automatique des Langues Naturelles, Senlis, France.

LESBEGUERIES, J. (2007). *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé.* Thèse de doctorat, Université de Pau et des Pays de l'Adour.

MARTINEAU C., TOLONE E., VOYATZI S. (2007). *Les Entités Nommées : usage et degrés de précision et de désambiguïsation*. 26ème Colloque international sur le Lexique et la Grammaire (LGC'07), Bonifacio, France. p. 105-112, 2007.

MAUREL Denis, FRIBURGER Nathalie, ANTOINE Jean-Yves, ESHKOL Iris, NOUVEL Damien (2011). *Cascades de transducteurs autour de la reconnaissance des entités nommées*. Traitement Automatique des Langues, ATALA, 2011, 52 (1), pp.69-96.

NADEAU N., SEKINE S. (2009). *A survey of named entity recognition and classification*. In S. Sekine & E. Ranchhod (eds.), John Benjamins publishing company, Amsterdam, pp. 3-28.

NOUVEL D. (2012). *Reconnaissance des entités nommées par exploration de règles d'annotation : Interpréter les marqueurs d'annotation comme instructions de structuration locale*. Thèse de doctorat, Université François Rabelais de Tours, Ecole Doctorale MIPTIS, Laboratoire d'Informatique, Equipe BdTln.

ROSSET S., GROUIN C., ZWEIGENBAUM P. (2011). *Entités Nommées Structurées : guide d'annotation Quaero*. Technical report.

De certains usages dans la twittosphère : contribution à une sociolinguistique computationnelle

Clément Thibert
Laboratoire ICAR - UMR 5191
CNRS, Université de Lyon & ENS de Lyon
clement.thibert@ens-lyon.fr

Résumé

Cette étude se propose de caractériser l'hétérogénéité et la variabilité des usages langagiers de communautés d'utilisateurs de Twitter. Nous abordons cette variabilité en examinant la distribution des parties du discours de l'ensemble des tweets de communautés, au sens de la science des réseaux, auxquelles des utilisateurs appartiennent. Les résultats mettent en évidence des usages qui diffèrent selon la visibilité des propos : d'un côté on trouve des communautés qui, donnant à voir leurs tweets, ont des usages proches des genres scripturaux, de l'autre des communautés qui, n'ayant pas ces pratiques, ont des usages plus proches des genres oraux.

Mots clés : sociolinguistique computationnelle, communication électronique médiée, variabilité linguistique, communauté, Twitter.

1 Introduction

L'interdisciplinarité est partout, en sciences du langage comme ailleurs en sciences et dans les domaines de l'ingénierie, où elle ne cesse d'augmenter depuis le milieu des années 80. Cette collaboration entre les disciplines est d'autant plus forte que la proportion d'articles référençant d'autres disciplines que la leur suit la courbe inverse de ceux citant des articles provenant exclusivement de leur propre discipline (Van Noorden, 2015). La sociolinguistique est un domaine qui n'échappe pas non plus à l'interdisciplinarité. Elle trouve d'ailleurs un essor particulier au sein de ce qu'on nomme comtemporainement la « sociolinguistique computationnelle » cherchant à résoudre des questions sociolinguistiques par l'assistance de moyens informatiques (voir Nguyen, (2015) pour une revue). Elle s'inscrit plus largement au sein des « sciences sociales computationnelles » qui illustrent aussi cette convergence plus large des sciences sociales avec, entre autres, l'informatique, le traitement automatique du langage naturel, la science des réseaux ou les statistiques (Lazer *et al*., 2009).

Cette contribution entend se situer à la jonction de la sociolinguistique, de la linguistique computationnelle, du traitement automatique du langage et de la science des réseaux. Elle étudie la notion de communauté, au sens de la science des réseaux, par une étude distributionnelle des parties du discours (désormais POS pour *parts-of-speech*) des tweets écrits par les utilisateurs qui composent ces communautés. Il s'agit plus précisément d'évaluer si les pratiques discursives sur Twitter sont homogènes ou si certaines communautés d'utilisateurs ont des pratiques discursives qui se rapprochent plutôt des genres oraux ou plutôt des genres écrits (Biber, 1988). Ces affinités pour l'oral ou l'écrit peuvent être abordées à travers la distribution des POS qui diffère selon le genre (Halliday, 1994 ; Biber *et al*., 1999)[1].

2 Vers une sociolinguistique computationnelle

2.1 Sociolinguistique des médias sociaux

La disponibilité sans précédent de données linguistiques et sociales, fait concomitant à l'explosion des usages des communications électroniques et à la génération massive de données issues de médias sociaux, a favorisé

[1] Notons, comme le fait Gadet (1996), que les observations d'Halliday (1994), valables pour l'anglais, le sont tout autant pour le français. La même remarque peut être étendue à Biber *et al*. (1999).

ce rapprochement entre la sociolinguistique et d'autres domaines comme la linguistique computationnelle et le traitement automatique du langage. Cette profusion de données, même si elle pose de nombreux défis et de nouveaux enjeux inhérents à leur provenance (Nguyen, 2015 ; Thibert, 2016), a provoqué un tournant inédit en faisant tomber le paradoxe de l'observateur formulé par Labov (1972). Il est en effet désormais facile de constituer des corpus de données spontanées/attestées et non biaisées par la présence de l'informateur.

Les approches computationnelles de la sociolinguistique ont particulièrement contribué à l'émergence, ces quinze dernières années, d'une littérature abondante ayant pour objet la communication électronique médiée (désormais CEM). Au même titre que d'autres formes de communications, la variabilité linguistique s'observe à travers toutes les formes de CEM. Les travaux de Paolillo (2001) ont montré, pour le tchat, une corrélation entre la variation linguistique et la position sociale des individus en sont les précurseurs. De même, dans les médias sociaux comme Twitter, l'âge, le genre ou la localisation géographique sont maintenant connus pour être des facteurs de variation (Bryden *et al*., 2013 ; Eisenstein, 2015 ; Goncalvez *et al*., 2015 ; Magué *et al*., 2015 ; Thibert *et al*., 2016).

2.2 Médias sociaux et science des réseaux

Dès lors qu'on s'intéresse aux médias sociaux, il semble naturel de recourir à la science des réseaux puisqu'elle étudie principalement les graphes, objets modélisant les interactions que des entités, organisées en réseaux, entretiennent entre elles. La notion de communauté est une importante propriété structurelle des réseaux qui désigne des entités densément connectées entre elles en des ensembles qui sont peu connectés les uns aux autres (Girvan & Newman, 2002). Cette notion a donné lieu à de nombreux travaux sur leur détection et leur caractérisation dans des domaines divers allant des réseaux de communication à la biologie, à la sociologie et aux neurosciences, entre autres (Malliaros & Vazirgiannis, 2013 ; Yang & Leskovec, 2015).

L'analyse des propriétés des réseaux des médias sociaux à travers cette notion de communauté et en relation à des questions linguistiques a fait apparaître plusieurs phénomènes tels que la dynamique des innovations (Altmann *et al*., 2011) et des emprunts (Garley & Hockenmaier, 2012 ; Eisenstein *et al*., 2014), ou encore la convergence linguistique (Danescu-Niculescu-Mizil *et al*., 2011 ; Tamburini *et al*., 2015). Il a notamment été montré que les communautés détectées sur Twitter étaient spatialisées et que leur structure était corrélée avec la distribution des fréquences lexicales (Magué *et al*., 2015).

3 Variation médiale et textométrie

La caractérisation des productions langagières en termes d'oralité et de scripturalité est un sujet pour lequel la littérature est abondante. Nous abordons ici principalement les spécificités distributionnelles des POS. Concernant la notion de genre, nous nous positionnons dans la lignée de Biber (1988) pour qui la pertinence de l'appartenance d'un texte à un genre est basée sur des critères linguistiques saillants mais également sur ce que l'on sait des intentions de l'auteur : « text categorizations made on the basis of external criteria relating to author/speaker purpose » (Biber, 1988 : 68).

Nous suivons ici le modèle de Koch & Oesterreicher (2001) d'après lequel les productions langagières peuvent être différenciées selon (i) une dichotomie tenant au code, phonique *vs* graphique, et selon (ii) un continuum conceptionnel, allant de l'immédiat communicatif (ayant des affinités pour l'oral) à la distance communicative (ayant des affinités pour l'écrit).

Avec ce modèle, Overbeck (2015) propose de placer les différents types de CEM selon le médium (oral/écrit) utilisé et selon le degré de proximité/distance conceptionnel. Ainsi, on trouve, dans cette classification, des CEM médiés par la phonie qui entretiennent des affinités (i) avec l'oral (comme la radiophonie) ou (ii) avec l'écrit (comme le blog vidéo) et ceux médiés pas la graphie qui entretiennent des affinités (iii) avec l'oral (comme le tchat) ou (iv) avec l'écrit (comme le courriel).

Par le passé, Halliday (1989) a formulé la distinction oral/écrit comme tenant à une « densité » singulière de chaque médium. L'écrit étant lexicalement dense, il comporte une forte proportion d'items lexicaux (pour une grande part, des noms). L'oral, quant à lui, est grammaticalement plus dense ; il comporte une forte proportion d'items grammaticaux.

Par ailleurs, il se caractérise également par davantage de verbes. Ainsi, dans le langage parlé, les mêmes phénomènes seront exprimés à l'oral par des verbes et à l'écrit par des nominalisations : « Written language tends to express phenomena like they were products whereas spoken language express phenomena as if they were processes » (Halliday, 1994 : 65).

De même, l'approche textométrique du genre textuel, développée plus tard par Biber *et al*. (1999), à travers l'examen de la distribution des POS, a mis en évidence des disparités entre les genres tenant plus de l'oral et ceux tenant plus de l'écrit. Le genre conversationnel est caractérisé par une haute fréquence de verbes et d'adverbes, une plus basse fréquence de noms et une forte proportion de pronoms, d'où une densité lexicale moindre. Les registres journalistique et académique sont, à l'opposé, caractérisés par une haute fréquence de noms, d'adjectifs, de déterminants et de prépositions. La densité lexicale y est par conséquent plus élevée (particulièrement pour les journaux). Par ailleurs, la distribution des conjonctions (coordonnants et subordonnants) n'a pas de préférence pour un genre.

Concernant les CEM, Panckhurst (2007) a étudié les distributions des POS dans le courriel, le forum et le tchat. Celles-ci se rapprochent de l'oral, avec cependant une proportion notable de verbes plus importante dans le courriel que dans les deux autres genres. À notre connaissance, une seule étude, portant sur le coréen, a comparé Twitter à d'autres genres. Son *et al*. (2014) ont évalué la distribution de 56 traits morphosyntaxiques en comparant un corpus de tweets (un ensemble de plus de 600 000 tweets) à 18 autres genres issus du Corpus du Coréen Standard (corpus comportant plus de 3 millions de mots). D'après leurs résultats, Twitter semble être un genre à part car il ne comporte aucune des spécificités qui caractérisent les autres genres étudiés, ni ceux proches de l'oral ni ceux proches de l'écrit, ceci suggérant que le tweet est un genre de type hybride.

Voici des exemples prototypiques issus du corpus de données qui illustrent cette tension oral/écrit où (1) et (2) relèvent plutôt de la scripturalité alors que les exemples (3) et (4) relèvent plutôt de l'oralité :

(1) Optimisation du lancement du launcher et autres actions en arrière plan :)

(2) Nous cherchons des traducteurs pour notre site, notre système de langue est déjà opérationnel.

(3) @mention1 @mention2 nan jconfirme les mecs ils ont trop de chance de nous avoir ! Ahahaha xD

(4) Ptn meme lmatin c la chaleur j'en peux plu jv crever 

On trouve en (1) et (2) une forte proportion de noms et de déterminants et, en outre, pour (2) une utilisation normée de la ponctuation. En revanche, en (3) et (4), on trouve une forte proportion de verbes et de pronoms et/ou d'adverbes. Notons qu'en dehors de ces considérations distributionnelles, ces tweets sont caractérisés par de nombreux traits typiques des CEM (Overbeck, 2015 ; Cougnon, 2016) : des smileys, des emojis, des interjections, des agglutinations et le non-marquage de la ponctuation.

4 Matériel et méthode

4.1 Acquisition des données

Le corpus de tweets initial est le résultat d'un échantillonnage réalisé sur une période d'environ un an (de juin 2014 à juin 2015) résultant d'une sélection aléatoire de 10% de la totalité des tweets (i) émis par des utilisateurs déclarant tweeter en français ou détectés comme français par Twitter (qui possède son propre détecteur de langue) et (ii) produits dans les fuseaux horaires GMT et GMT+1 dans ce que l'on peut nommer les espaces francophones européen et africain. Les profils des utilisateurs ainsi que la liste des *followers* de chaque utilisateur dont au moins un tweet est présent dans le corpus ont également été récupérés. Les données résultant de cet échantillonnage sont composées de près de 70 millions de tweets et d'un réseau constitué d'environ 1,7 million d'utilisateurs.

4.2 Reconstruction du réseau, détection et filtrage des communautés

La liste des utilisateurs et la liste des *followers* de chaque utilisateur ont servi de base à la reconstruction du réseau de relations. Ceci nous a permis de déterminer si les relations entre les utilisateurs sont unidirectionnelles (un utilisateur suit un autre utilisateur sans que ce dernier le suive) ou réciproques (deux utilisateurs se suivent l'un l'autre, selon le principe *follower-followee*). Les utilisateurs entretenant des relations unidirectionnelles avec d'autres

utilisateurs ont été écartés car nous avons considéré qu'il n'existe pas de relation de proximité assez forte entre deux utilisateurs qui ne se suivent pas l'un l'autre. Nous avons ensuite utilisé un algorithme de détection de communautés, celui de Louvain (Blondel *et al*., 2008), afin de déterminer si les utilisateurs du réseau appartiennent à des sous-ensembles densément connectés. Nous avons finalement filtré les communautés en ne conservant que celles comportant au moins 1000 utilisateurs et dont au moins 50 % des tweets ont été détectés comme français par le détecteur Ldig (Lui & Baldwin, 2014). De ce filtrage résulte un corpus de 57 122 195 tweets produits par 701 791 utilisateurs répartis à travers 14 communautés. Le tableau ci-dessous rapporte le nombre d'utilisateurs, le nombre de tweets et la part (en %) de tweets français pour chaque communauté associée à un identifiant.

Com. id	Utilisateurs	Tweets	% en fr.
1	334 614	38 189 807	62,34
2	299 594	12 686 580	76,67
3	29 856	3 820 909	60,47
4	9 993	769 168	57,95
5	8 109	485 460	73,68
6	3 746	333 546	71,18
7	3 685	192 513	72,48
8	2 392	108 555	75,44
9	2 299	125 629	73,61
10	1 801	120 420	64,78
11	1 593	27 687	52,70
12	1 541	128 680	69,22
13	1 366	16 135	46,18
14	1 202	117 106	69,38
Total	701 791	57 122 195	moy. = 65,59

Tableau 1 : répartition du nombre d'utilisateurs, de tweets et part des tweets en français par communauté

4.3 Annotation morphosyntaxique et calcul des fréquences

L'ensemble des tweets du corpus ont été annotés avec MElt (Denis & Sagot, 2009), annotateur morphosyntaxique spécialisé dans le traitement des textes « bruités », tels que ceux produits dans les médias sociaux, et entraîné sur le *French Social Media Bank* (Seddah *et al*., 2012). La version que nous avons utilisée est une version adaptée au traitement des tweets. La liste des POS et de leur étiquette associée est la suivante: adjectif (Adj), adverbe (Adv), conjonction (Cnj), déterminant (Dét), interjection (Int), nom commun (NomC), nom propre (NomP), pronom (Pro), préposition (Pré) et verbe (Ver). À ces étiquettes s'ajoutent la ponctuation (Pct), les mots inconnus (Inc), c'est-à-dire les tokens non reconnus par MElt, et les emoji (Emo). Par souci de simplification, nous avons également rangé dans les POS les éléments de ces quatre dernières sortes. Les fréquences cumulées de chacune des POS ont été calculées pour chaque communauté. Nous avons également dénombré les hashtags, les mentions, les URL et les tokens.

4.4 Analyses

Afin de déterminer l'existence de liens entre les variables, nous avons procédé à une analyse factorielle des correspondances (désormais AFC). Nous avons aussi procédé à une classification hiérarchique sur composantes principales (désormais CHCP) pour déterminer si certaines communautés avaient des profils similaires et comment celles-ci se regroupaient. Les analyses ont été faites à l'aide du package FactoMineR (Lê *et al*., 2008).

5 Résultats

5.1 Analyse factorielle des correspondances

La figure 1 présente la projection des communautés et des POS sur les deux premières dimensions de l'AFC. Ces deux premières dimensions contiennent 98,01% de l'inertie totale : la première dimension (l'axe horizontal) permet d'expliquer 94,70% de l'inertie et la seconde (l'axe vertical) en explique 3,31%. Le premier axe résume bien à lui seul l'écart à l'indépendance[2] ce qui nous permet de nous limiter à celui-ci pour l'interprétation des résultats. Deux communautés (1 et 2) contribuent à elles seules à plus de 99% de la construction des axes. Concernant la qualité de projection, 11 des 14 communautés sont relativement bien projetées ($cos^2 > 0.54$; moy. = 0.73). Les communautés 3, 4 et 11 (pour qui les cos^2 sont proches de zéro) ne sont pas prises en compte dans la suite des analyses. Quasiment la totalité des POS est bien projetée ($cos^2 > 0.79$; moy. = 0.93) ; la catégorie des mentions, très moyennement projetée ($cos^2 = 0.48$) ainsi que les catégories des interjections et des mots inconnus (pour qui les cos^2 sont proches de zéro) ne sont pas prises en compte dans la suite des analyses.

[2] L'écart à l'indépendance représente ici la différence entre les effectifs observés et les effectifs théoriques des fréquences. L'effectif théorique est l'effectif que l'on observerait si les deux modalités (fréquences et communautés) était indépendantes.

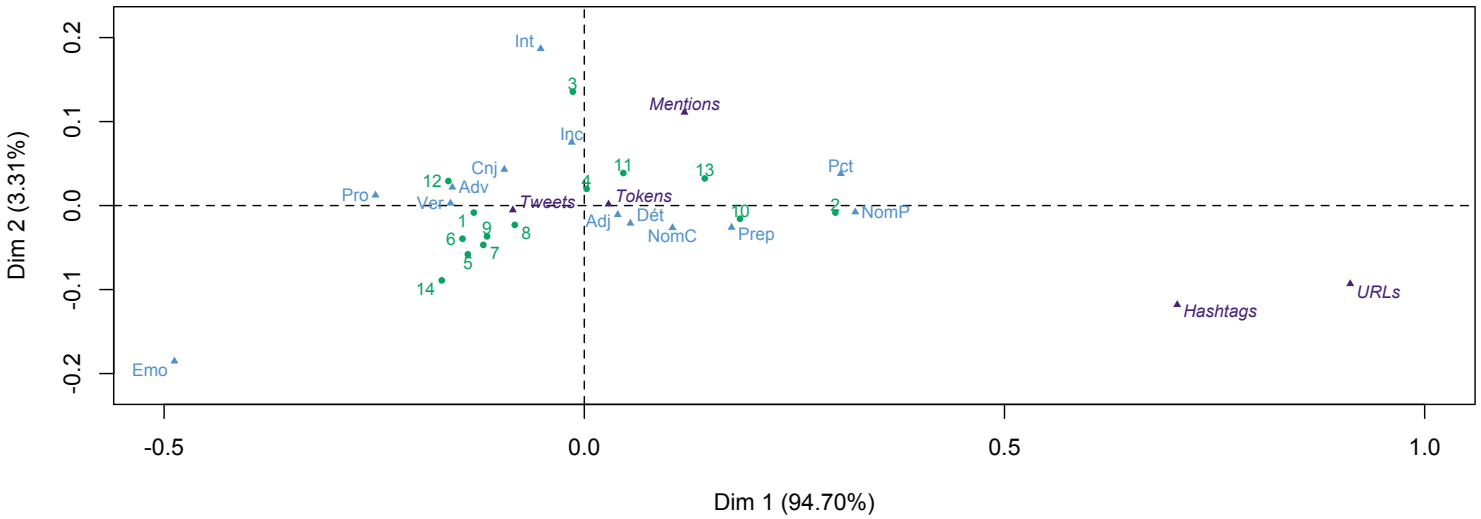


Figure 1 – Analyse factorielle des correspondances

La première dimension (le long de l'axe x) semble opposer deux ensembles de communautés selon la répartition des POS. Un premier ensemble (situé à gauche de l'axe) utilise principalement des verbes, des adverbes, des pronoms, des conjonctions et des emojis et produit également une plus grande quantité de tweets que les autres communautés. Cet ensemble est composé des communautés 1, 5, 6, 7, 8, 9, 12 et 14. Un second ensemble (situé à droite de l'axe) utilise plutôt des déterminants, des noms communs, des adjectifs, des noms propres, des prépositions et des signes de ponctuation. Cet ensemble est également caractérisé par une utilisation plus forte des mentions, des hashtags et du partage d'URL et par une longueur de tweets plus importante par rapport aux autres communautés. Cet ensemble est composé des communautés 2, 10 et 13.

5.2 Classification hiérarchique sur composantes principales

La figure 2 rapporte les résultats de la CHCP. L'arbre hiérarchique suggère une partition des communautés en 4 ensembles distincts. Les deux premiers ensembles correspondent à ceux qui ont été mis en relief par les résultats de l'AFC. Le premier présenté (cerné de bleu) est composé des communautés 2, 10 et 13. Le deuxième (cerné de noir) est composé des communautés 1, 5, 6, 7, 8, 9, 12 et 14. Les deux autres ensembles, la communauté 4 (cerné de vert) d'une part et les communautés 3 et 11 (cerné de rouge) d'autre part, correspondent aux communautés que l'on ne pouvait pas prendre en compte lors des analyses car elles sont mal représentées sur l'AFC.

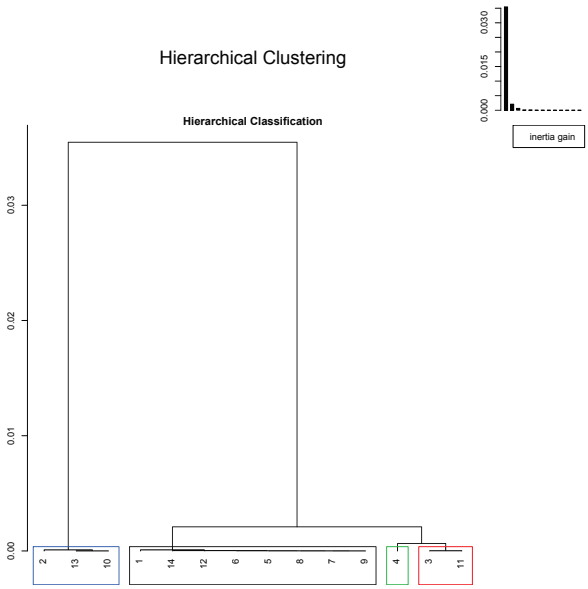


Figure 2 – Classification hiérarchique des communautés

6 Discussion

L'approche textométrique adoptée ici, à l'instar de Biber *et al*. (1999), a permis de caractériser les communautés ayant des pratiques langagières qui diffèrent selon leurs affinités avec des genres relevant plutôt de l'oral ou au contraire plutôt de l'écrit. Les résultats présentés dans nos analyses sont concordants avec les observations et les analyses d'Halliday (1994), Biber *et al.* (1999) et Panckhurst (2007) susmentionnées en ce qui concerne les affinités de genres selon la distribution des parties du discours. Les pratiques langagières des communautés d'utilisateurs de Twitter observées ici sont disparates: la répartition des fréquences

des POS étant distribuée non uniformément à travers les communautés.

Les résultats présentés montrent que Twitter, à l'échelle des communautés, ne forme pas un tout homogène, faisant genre en soi, mais est composé d'une multiplicité de genres qui coexistent. L'AFC a mis en exergue des préférences pour des types de production ayant des affinités soit avec les genres oraux soit avec les genres écrits qui semble dépendre des situations de communication. D'un côté, on trouve des communautés d'utilisateurs partageant du contenu via des URL et indexant leurs propos avec des hashtags, les utilisateurs donnant ainsi à voir leurs tweets à une audience large. S'exprimant dans la sphère publique, les utilisateurs appartenant à ces communautés utilisent une langue à dominante scripturale car caractérisée, entre autres, par une prédominance de noms, de déterminants et d'adjectifs et par une forte utilisation de la ponctuation. A l'opposé, de ces communautés d'utilisateurs, on trouve d'autres communautés n'ayant pas ces pratiques de diffusion et de partage. Ces utilisateurs paraissent recourir à une langue à dominante orale car caractérisée, entre autres, par une prédominance de verbes, d'adverbes et de pronoms. Notons néanmoins que l'on aurait pu s'attendre à trouver les mentions allant dans ce sens, ce qui n'est pas le cas. Il semble difficile d'interpréter cette donnée, en l'état, mais l'on peut songer au fait que les utilisateurs de Twitter s'adressent sans doute différemment à leurs *followers* selon le type de relation qu'ils entretiennent et qu'il sera nécessaire de prendre en compte cette information dans les prochaines analyses.

De nombreuses perspectives sont envisageables pour prolonger ces observations. Même si la notion de communauté, au sens de la science des réseaux, fait sens, nous n'avons pas, pour le moment, une connaissance fine des utilisateurs qui les composent. Connaître leurs caractéristiques sociodémographiques, entre autres, pourrait nous permettre de mettre ces informations plus finement en relation avec les affinités de genres dégagées ici. En lien avec cela, et au-delà d'une seule caractérisation des communautés de scripteurs de Twitter à partir de l'analyse de la distribution des POS, cette méthode mérite d'être réutilisée en s'intéressant aux diverses variantes lexicales ou syntaxiques potentiellement identifiables dans une perspective issue de la sociolinguistique variationniste.

Au-delà des affinités pour les genres oraux ou pour les genres scripturaux, démontrées ici à l'échelle des communautés, on peut aussi aisément supposer que les utilisateurs ne sont pas nécessairement cantonnés à un usage plutôt qu'à un autre mais qu'ils adaptent leurs pratiques discursives, celles-ci n'étant sans doute pas uniformes à travers leurs échanges sur Twitter mais dépendantes de la situation de communication et/ou des relations que les scripteurs entretiennent entre eux. Cette variation probable entre des tweets donnés à voir à la « twittosphère » et des tweets plus « privés » pourra être examinée plus précisément. La mise en évidence d'une adaptation des usages des utilisateurs au contexte pourra aussi être envisagée, à la suite de Cougnon (2016), en lien avec la compétence/capacité des individus à jouer avec les variations et la norme.

7 Conclusion

L'hétérogénéité et la variabilité des usages langagiers de communautés d'utilisateurs de Twitter ont été abordées ici par une approche à l'intersection de la sociolinguistique, du traitement automatique du langage et de la science des réseaux. Bien qu'il s'agisse d'une étude exploratoire, la robustesse de la méthode utilisée et la pertinence des résultats concourent à faire de la sociolinguistique computationnelle un champ de recherche plein de promesses pour étudier, à la fois à grande échelle et qualitativement, la variation sociolinguistique des usages sur les médias sociaux.

Remerciements

Cette contribution a reçu le soutien financier de l'Agence Nationale de la Recherche à travers le projet SoSweet (ANR-15-CE38-0011-01) et à travers le LabEx ASLAN, Laboratoire d'Excellence des études avancées sur la complexité du langage (ANR-10-LABX-0081).

Références

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PloS one*, 6(5), e19009.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman*

Grammar of Spoken and Written English, (2). MIT Press.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, (10), P10008.

Bryden, J., Funk, S., & Jansen, V. A. (2013). Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*, 2(1), 3.

Cougnon, L.-A. (2016). « Conflit, réinvention et variation de normes de communication dans la CMO », dans Gaudin-Bordes, L. & Monte, M., (dirs), *Normes textuelles : émergence, variations, conflits*.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, 745-754.

Denis, P. & Sagot, B. (2009). Coupling an annotated corpus and a lexicon for state-of-the art P.O.S. tagging. *Language Resources and Evaluation.* 46(4), 721-736.

Eisenstein, J. (2015). Written dialect variation in online social media in Boberg, C., Nerbonne, J.& Watt, D., (eds), *Handbook of Dialectology* (Wiley).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, 9(11), e113114.

Gadet, F. (1996). Une distinction bien fragile : oral/écrit, *Tranel*, 25, 13-27.

Garley, M., & Hockenmaier, J. (2012, July). Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers,* 2, 135-139.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing dialect characterization through Twitter. *PloS one*, *9*(11), e112074.

Halliday, M. A. K. (1994). Spoken and Written Modes of Meaning. *Media texts, authors and readers*, 51-73.

Halliday, M. A. K. (1989). *Spoken and Written Language*. Geelong, Victoria: Deakin University Press (republished by Oxford University Press in 1989).

Koch, P. & Oesterreicher, W. (2001). « Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit », *Lexicon der Romanistischen Linguistik*, 1 (2), Niemeyer, Tübingen, 584-627.

Labov, W. (1972). Language in the inner city: Studies in the Black English vernacular, 3. University of Pennsylvania Press.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*. 25(1), 1-18.

Lui, M. & Baldwin, T. (2014). Accurate language identification of twitter messages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Göteborg, Sweden, April. ACL. 5th Workshop on Language Analysis for Social Media., 17-25.

Magué, J.-P., Fleury, E., Karsai, M. & Quignard, M. (2015). Caractérisation dialectale de la variabilité linguistique sur Twitter. Language, Cognition and Society (AFLiCo6), Grenoble, Mai.

Malliaros, F. D., & Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4), 95-142.

Nguyen, D., Rosé, C.P., Seza Dogruöz, A. & de Jong, F. (2015). Computational Sociolinguistics : A Survey. arXiv: 1508.07544v1.

Overbeck, A. (2015). La communication dans les médias électroniques. *Manuel de linguistique française*, 8, 275-292.

Panckhurst, R. (2007). Discours électronique médié : quelle évolution depuis une décennie ? In Gerbault, J. (éd.) *La langue du cyberspace : de la diversité aux normes*, L'Harmattan, Paris, 121-136.

Paolillo, J. C. (2001). Language variation on Internet Relay Chat: A social network approach. *Journal of sociolinguistics*, 5(2), 180-213.

Seddah, D., Sagot, B., Candito, M, Ouilleron, V. & Combet, V. (2012). ‘‘The French Social Media Bank: a Treebank of Noisy User Generated Content”. In Kay, M. & Boitet, C. (Ed.). *Proceedings of CoLing 2012: Technical Papers*, 8-15 Decembre 2012, Mumbai, India, 2441-2458.

Son H., Lee, J.Y., Kang, B. & Kim H (2014). Twitter en coréen: un langage d'un genre nouveau. *Faits de langues.* Varia, (41), 125-144.

Tamburrini, N., Cinnirella, M., Jansen, V. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40, 84-89.

Thibert, C. (2016). Twitter as Corpus for Sociolinguistics Variationnist Studies: Challenges of Using Sketchy Data. Workshop: Using Twitter for Linguistic Research. Canterbury, University of Kent. May, 31.

Thibert, C., Magué, J.-P., Fleury, E., Karsai, M. & Quignard M. (2016). Dialectal Characterization of Linguistics Variability on Twitter. Data Driven Approaches to Networks and Language. Lyon. May, 11-13.

Van Noorden, R. (2015). Interdisciplinarity Research by the numbers. *Nature,* 525(7569). 306-307.

Yang, J. & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*. 42(1).

Méthode hybride pour l'identification automatique de la langue sur textes courts et très courts

Valentin NYZAM
valentin.nyzam@gmail.com
Mohamed Slim BEN MAHMOUD
mohamedslim.benmahmoud@altran.com

Abstract

Dans le cadre d'études en traitement automatique du langage, il est primordial de pouvoir identifier de manière précise la langue du texte étudié. Si ce problème est considéré comme résolu pour des textes de la taille d'une phrase (à partir de 5 mots environ) ou d'un tweet dans les dernières recherches, les résultats sont beaucoup moins probants pour des textes plus courts voire de la taille d'un simple mot. Dans cette étude, nous allons étudier les différents types de méthodes qui ont été mis en place jusqu'à aujourd'hui dans ce contexte, avant de présenter une nouvelle méthode hybride ayant pour objectif d'améliorer les résultats existants.

MOTS-CLÉS : *Identification Automatique de la Langue, Textes très courts, Méthode hybride.*

1 Introduction et problématique

L'identification automatique de la langue d'un texte donné est indispensable pour de nombreuses applications. Elle est souvent la première étape de traitement d'un système informatique utilisant la langue naturelle. De tels systèmes appliquent en effet des chaînes de traitement utilisant des modèles de langues. Si ce problème est considéré comme résolu d'après (McNamee, 2005), cela est vrai sur un texte donné de taille suffisante. Pour des textes courts voire très courts (en dessous 300 caractères environ d'après (Tromp and Pechenizkiy, 2011), les résultats se dégradent rapidement et peu de travaux se sont intéressés à la précision des algorithmes sur un simple mot.

L'identification statistique de mots, sans passer par la constitution de dictionnaire complet de langues, est problématique du fait de la redondance de certains mots dans plusieurs langues (*e.g.* place en français/anglais, botte en français/italien) parfois avec un sens différent. Il est ainsi primordial de pouvoir classer ces mots comme un cas d'indétermination.

Nous allons donc, dans ce papier, commencer par présenter les critères d'identification possibles sur lesquels les méthodes existantes s'appuient. Dans une seconde partie, nous testerons quelques méthodes représentatives de l'état de l'art ainsi que leurs résultats sur des textes très courts. En dernier lieu, nous présenterons le développement d'une méthode hybride améliorant alors les résultats précédents.

2 Linguistique et critères d'identification d'une langue

Un algorithme d'identification de la langue est capable de prédire automatiquement la langue d'un texte donné. Quand une personne cherche à identifier manuellement la langue d'un texte, elle se base habituellement sur les caractères uniques ou typiques de certaines langues. Cela peut être une suite de lettres communes ou particulières, le début ou la fin des mots (préfixe et suffixe), ou les mots grammaticaux (appelés aussi mots-outils). Ces critères, par leurs présences ou leurs absences, sont des indices forts pour l'identification.

D'autre part, les algorithmes d'identification sont habituellement basés sur des dictionnaires ou des méthodes statistiques, voire une combinaison des deux.

Les méthodes par dictionnaire se basent sur des listes de mots spécifiques à chaque langue. Les mots composant une langue peuvent en effet être séparés en deux grandes catégories :

1. Les mots lexicaux qui correspondent aux noms, verbes, adjectifs qualificatifs et adverbes. Ils sont très nombreux et en création continue. D'un point de vue sémantique, les mots lexicaux sont le plus souvent susceptibles d'avoir plusieurs sens mais dans un

contexte donné, chacun d'entre eux constitue une unité de sens.

2. Les mots grammaticaux qui correspondent aux déterminants, pronoms et conjonctions. Généralement courts, les mots grammaticaux sont en nombre limité et il est plus facile d'en dresser des listes par comparaison aux mots lexicaux, qui sont eux plus nombreux. Les mots grammaticaux ont un caractère obligatoire et un rôle plus syntaxique que sémantique. Le pronom est un cas particulier des mots grammaticaux car ils sont souvent présentés comme étant à cheval entre les unités lexicales et les unités grammaticales (peu nombreux, ils entrent dans des relations syntaxiques variées en remplaçant un nom, un groupe nominal, un adjectif ou une proposition). On parle alors de substitut car il peut remplacer autre chose que le nom. Ainsi, il a un rôle syntaxique important et ce type de mots grammaticaux devrait être les plus rencontrés dans des textes très courts. Nous avons fait le choix de les classer comme mots grammaticaux pour simplifier l'étude.

En établissant une liste de ces mots grammaticaux pour plusieurs langues, il est alors possible d'utiliser ces mots comme caractère discriminant pour l'identification de la langue. Néanmoins, l'orthographe de ces mots se retrouve souvent dans plusieurs langues. Par exemple, le mot "a" se retrouve en anglais, en français et en espagnol. Afin de répondre à ce problème, (Rehurek and Kolkus, 2009) ont mis en place un algorithme d'identification basé sur une fonction de pertinence. À l'aide d'une méthode d'apprentissage automatique sur de grands corpus, ils utilisent une méthode proche de TF-IDF (Term Frequency-Inverse Document Frequency) afin de déterminer un score de pertinence par langue pour chaque mot.

La difficulté et le coût des méthodes par dictionnaire a plutôt mis en avant le développement des méthodes statistiques. Ces méthodes se basent sur l'apprentissage des suites de caractères les plus fréquentes d'une langue.

En effet, chaque langue a un vocabulaire propre mais peut aussi avoir une racine (ou base linguistique) commune avec une ou plusieurs langues. En général, les langues n'utilisent pas les mêmes voyelles tout en possédant une même racine. De même, certaines langues utilisent des alphabets qui leurs sont propres. Par exemple, les voyelles "i" et "o" sont très utilisées en italien alors qu'en français ce sont les voyelles "e" et "a" qui sont les plus fréquentes, malgré une base linguistique commune et très proche. Chaque langue possède aussi des N-grammes (Suite de N caractères) caractéristiques. Par exemple, le 2-gramme "de" est très utilisé en français alors que le 3-gramme "ing" l'est en inglais. Cette idée est représentée par la loi de Zipf qui énonce que : "le N-ième mot le plus commun d'un texte en langue naturelle apparaît avec une fréquence inversement proportionnelle à N". Ce sont ces trois critères principaux qui sont utilisés par les méthodes statistiques même si l'approche la plus usitée utilise les N-grammes. Il a été montré qu'elle a quasiment 100% de précision sur des textes suffisament long. L'architecture globale d'un système d'identification de la langue a ainsi été donné par (Padró and Padró, 2004) dans la figure 1.

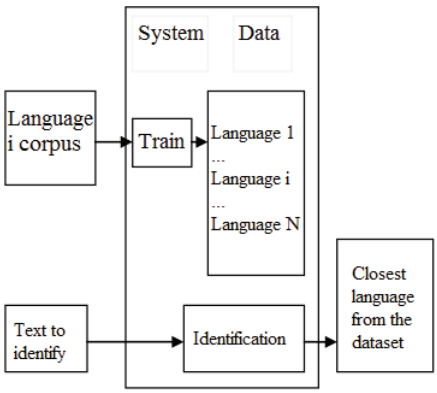


FIGURE 1 – Architecture générale d'un système statistique d'identification de la langue.

Un algorithme d'identification de la langue peut alors être caractérisé selon différents critères :

— Les modèles de langue, représentés par les probabilités d'apparition des N-grammes (ou des mots) pour chaque langue,
— La taille des entités à reconnaître (texte, phrase, groupe de mots, un mot),
— Les caractéristiques principales de l'algorithme (statistique ou basé sur des règles ; basé sur les caractères ou sur les mots),
— La précision et la validité des résultats,
— La complexité et la vitesse d'exécution,
— Le nombre de langues à traiter,
— La gestion des langues inconnues, et de l'indétermination.

Ainsi, une méthode globale d'identification se

résume comme suit : à l'aide d'un algorithme d'apprentissage, nous générons un profil (ou modèle) de langue pour chaque langue que nous souhaitons identifier. Ensuite, nous comparons le document à identifier avec les modèles de langue obtenus précédemment à l'aide d'un classifieur. Ce classifieur permet de calculer un "score" pour chaque langue, le score le plus haut correspond alors à la langue la plus probable du document.

Dans la suite de cette article, nous présenterons succinctement les méthodes et les classifieurs qui nous ont servis de référence pour notre propre méthode.

3 État de l'art des classifieurs existants

Notre objectif étant d'étudier les résultats d'identification de la langue sur des textes très courts (voire des mots), nous avons décidé de nous intéresser en premier lieu aux méthodes statistiques, qui sont les plus rapides à mettre en oeuvre (basées sur les N-grammes), et donc les plus utilisées. Un des premiers classifieurs mis en place pour l'identification de la langue est le classifieur "Out of Place" de (Cavnar and Trenkle, 1994). Celui-ci utilise des modèles de langue, générés par apprentissage automatique, composés de la liste des 100 à 400 premiers N-grammes (pour N variant de 1 à 5) de la langue, et triés par fréquence (par exemple, en français, le premier 2-gramme est "de"). Cavnar génère ensuite un modèle pour le texte de test. Le classifieur compare alors les positions des N-grammes de ce modèle avec les positions dans les modèles de chaque langue. La différence de position entre les modèles de test et de langue lui donne alors un score. Le score le plus faible (plus la différence de position est faible, plus le modèle de langue est "proche" du modèle de test) parmi les modèles des différentes langues correspond alors à la langue la plus probable.

(Dunning, 1994) a mis en place un nouveau classifieur utilisant les probabilités. À partir du nombre d'occurrence de chaque N-gramme obtenu lors de l'apprentissage, il génère une distribution de probabilité pour chaque N-gramme. Pour cette génération, Dunning utilise les modèles de Markov (avec S une chaîne composée des caractères $s_1 \ldots s_n$ et A le modèle de Markov) comme suit :

$$p(S|A)=p(s_1 \ldots s_n|A)=p(s_1|A) \prod_{i=2}^{n} p(s_i|s_{i-1}|A) \quad (1)$$

Le classifieur s'appuie ensuite sur le théorème de Bayes afin d'obtenir pour le texte de test les probabilités d'appartenance à chaque langue (avec un évènements A étant donné une observation X) :

$$p(A,X)=p(A|X)\cdot p(X)=p(X|A)\cdot p(A) \quad (2)$$

La formule globale devient alors (avec S une chaîne composé des caractères $s_1 \ldots s_n$ et un modèle de Markov A généré sur une langue à identifier) :

$$p(S|A)=p(s_1 \ldots s_k|A) \prod_{i=k+1}^{n} p(s_i|s_{i-k} \ldots s_{i-1}|A) \quad (3)$$

En calculant cette probabilité $p(S)$ pour chacune des langues apprises par le système, la langue la plus probable correspond à celle qui a la plus haute probabilité d'appartenance.

Afin de maximiser nos tests sur des textes très courts, nous souhaitons aussi apréhender l'utilisation de méthode par dictionnaire. Pour cela, nous avons décidé de mettre en place la méthode de (Giguet, 1995). Celui-ci a proposé une nouvelle méthode utilisant un classifieur basé sur les mots grammaticaux et sur les N-grammes. Son objectif était d'obtenir de bons résultats sur des textes bruités (obtenus par reconnaissance optique de caractères) ou comprenant des mots étrangers. La liste des mots grammaticaux est construite manuellement par Giguet pour quatres langues et est composée d'en moyenne deux cents mots grammaticaux. Il applique tout d'abord une reconnaissance des mots grammaticaux du texte à identifier, obtenant un premier score. Il utilise ensuite le classifieur de Cavnar & Trenkle afin d'obtenir un second score qu'il somme avec le précedent. En combinant ces deux méthodes, Giguet obtient un résultat plus efficace sur les phrases relativement longues bruitées, mais cette méthode est moins efficace sur les phrases courtes, du fait du manque de mots grammaticaux dans ce contexte.

(Teahan, 2000) a développé un classifieur innovant nommé PPM (Prediction by Partial Match). Celui-ci, bien qu'utilisant lui aussi les modèles de Markov, se base sur les travaux de la théorie de l'information de Shannon. Dans la théorie de l'information, le théorème de codage fondamental indique que la borne inférieure du nombre moyen de bits par symbole nécessaire à encoder un message est donnée par son entropie (avec P la distribution de probabilité d'un message composé de k symboles appartenant à un alphabet A) :

$$H(P)=-\sum_{i=1}^{k} p(x_i)\cdot \log p(x_i) \quad (4)$$

Cette formule peut ensuite être généralisée pour un langage ayant une distribution de probabilité L :

$$H(L) = \lim_{m \to \infty} -\frac{1}{m} \sum p(x_1,\ldots,x_m) \cdot \log p(x_1,\ldots,x_m) \quad (5)$$

$H(L)$ est alors l'entropie du langage et peut être considérée comme la limite de l'entropie quand le message devient très grand. Usuellement, la véritable distribution de probabilité L n'est pas connue. Toutefois, une borne haute à $H(L)$ peut être obtenue en utilisant un modèle M comme une approximation du langage L :

$$H(L,M) = -\sum p_M(x_1,\ldots,x_m) \cdot \log p_M(x_1,\ldots,x_m) \quad (6)$$

$H(L, M)$ est alors appelée l'entropie croisée (ou cross-entropy) et est toujours supérieure ou égale à $H(L)$. Le calcul de l'entropie croisée permet ainsi de mesurer la manière dont se comporte le modèle M par rapport à un texte de test : plus sa valeur sera proche de $H(L)$, plus le modèle sera exact. Ainsi, cela va permettre de comparer la précision de différents modèles. En calculant les entropies croisées pour chaque modèle de langue (*i.e.* pour la distribution de probabilité de chaque langue obtenue à l'aide d'un modèle de Markov), le modèle le plus proche de notre texte de test est celui qui obtient la valeur d'entropie croisée la plus faible.

D'autres classifieurs existent comme le classifieur SVM (Support Vector Machine (Hsu et al., 2003)) ou le classifieur Cosine Similarity (Brown, 2013) mais n'ont pas encore été testés dans nos travaux. Dans la section suivante, nous nous attarderons sur les corpus choisis dans notre étude.

3.1 Corpus d'apprentissage et de test

Afin de comparer les résultats sur une base commune, nous avons choisi comme corpus d'apprentissage des textes, libres de droit et issus du parlement Européen (Europarl Corpus of European Parliament Proceedings ou EPP [1]). Ces textes sont disponibles dans vingt et une langues et sont très volumineux (composés de plusieurs dizaines de millions de mots). Cela a permis par la suite d'étudier l'influence de la taille du corpus d'apprentissage sur les résultats d'identification. Nous avons ainsi déterminé qu'entre environ cinquante mille et cent mille caractères, les résultats d'identification atteignent tous une asymptote. Au-delà de cette valeur, augmenter la taille du corpus n'a plus aucune influence, si ce n'est un très léger sur-apprentissage.

Lors de l'apprentissage, il faudra néanmoins ne pas prendre en compte les noms propres, car cela fausserait les probabilités d'occurrence de chaque N-gramme. Pour cela, nous appliquons un prétraitement simple qui retire les mots commençant par une majuscule. Pour cela, l'allemand n'a pas pu être pris en compte, cette langue utilisant massivement les majuscules pour des mots autre que des noms propres.

Le corpus de test quant à lui est fabriqué manuellement. En effet, n'ayant trouvé aucun corpus composé de textes très courts (de un à cinq mots), nous avons récupéré des romans libres de droit disponibles sur le projet Gutenberg [2] que nous avons découpés en textes très courts d'un seul mot de cinq à quatorze caractères afin d'obtenir environ quatre milles échantillons de tests dans chaque cas.

Dans la suite, nous présenterons les résultats des méthodes de Cavnar & Trenkle, Dunning, Giguet et Teahan (utilisant les classifieurs "Out Of Place", Bayes, et PPM) sous les conditions de test du tableau suivant :

Langues prises en compte	français, anglais, dannois, finnois, portugais, espagnol, italien
Corpus d'apprentissage	Issue de EPP, comportant 100 000 caractères pour chaque langue
Corpus de test	Composé d'en moyenne 4000 mots de 7 caractères extraits de romans libres de droit pour chaque langue
Modèles de langues	Composés des 700 N-grammes les plus fréquents pour chaque langue
Cavnar & Trenkle	N-grammes pour N variant de 2 à 4
Dunning	Modèles de Markov d'ordre 2 à 4
Teahan	Modèles de Markov d'ordre 2 à 4

TABLE 1 – Conditions de test utilisées pour les expériences.

Les critères de mesure des performances seront le rappel et la précision à travers leur moyenne harmonique nommée F-mesure :

$$\text{Précision}_i = \frac{\text{Nb de documents correctement attribués à la langue } i}{\text{Nb de documents attribués à la langue } i} \quad (7)$$

$$\text{Rappel}_i = \frac{\text{Nb de documents correctement attribués à la langue } i}{\text{Nb de documents appartenant à la langue } i} \quad (8)$$

$$\text{F-Mesure} = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (9)$$

1. http://www.statmt.org/europarl/archives.html

2. https://www.gutenberg.org

3.2 Résultats des classifieurs existants

Comme précisé dans la section 1, il est primordial de pouvoir identifier les cas indéterminés tout en minimisant le nombre d'erreurs. Pour cela, nous mettons en place un seuil d'indétermination permettant d'obtenir un résulat d'indéterminé lorsque deux langues obtiennent des scores trop proches. Pour cela, nous faisons en sorte que les méthodes obtiennent un résultat indéterminé lorsque la différence de scores entre les deux meilleures langues est en dessous d'une certaine valeur obtenue grâce aux résultats de la figure 2. Nous fixons la valeur seuil à 0.2 car c'est celle qui offre le meilleur compromis entre les erreurs d'identification et le taux d'identification positive. Il est néanmoins possible de faire varier cette valeur en fonction du résultat recherché (*e.g.* maximisation du taux de réussite, minimisation du taux d'erreurs).

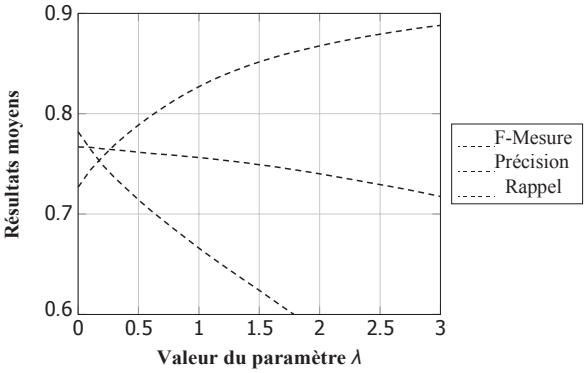


FIGURE 2 – Variation des résultats pour la méthode de Teahan en fonction de la valeur du seuil d'indétermination.

Comme on peut le voir dans le tableau 2, les résultats des classifieurs dépendent fortement des langues à identifier et donc, du pool de langues choisies. En effet, le portugais et l'espagnol étant des langues relativement proches linguistiquement, les identifications sur ces deux langues sont plus difficiles, de même pour le français et l'italien. En testant avec le même pool de langue mais en retirant l'espagnol, on obtient alors une précision de 81.41% pour le portugais avec la méthode de Teahan, ce qui montre bien les corrélations entre ces deux langues.

Ensuite, la méthode de (Cavnar and Trenkle, 1994) apporte une base d'identification intéressante mais qui reste faible avec une moyenne de 52.54% pour la F-Mesure. L'utilisation d'un classifieur plus abouti (utilisant les statistiques et le théorème de Bayes) avec la méthode de (Dunning, 1994) apporte une légère amélioration avec seulement 2.46% d'augmentation en moyenne par rapport à Cavnar & Trenkle. Cette faible différence est due à l'identification sur un nombre très faible de mots (pour rappel, deux mots). Sur des textes de longueur moyenne (entre six et vingt mots), les résultats du classifieur de Dunning sont meilleurs. Les scores obtenus par la méthode de Giguet sont plus faibles que ceux de la méthode de Cavnar & Trenkle. Cela est dû au très important taux d'erreurs et au taux d'indétermination plus faible engendré par l'identification par mots grammaticaux. En effet, l'identification par dictionnaire engendre une perte d'information car ceux-ci sont organisés en liste ne possédant pas de probabilité d'occurence contrairement aux modèles de N-grammes. De plus, la méthode de Giguet est très dépendante des listes de mots grammaticaux construites manuellement. Notre corpus de test étant composé de mots seuls de sept caractères, il ne comprend donc que peu de mots grammaticaux, ce qui doit diminuer les performances.

Le classifieur PPM utilisé par (Teahan, 2000), apporte les meilleurs résultats avec une amélioration moyenne de 22.39% par rapport à la méthode de Cavnar & Trenkle et 19.93% par rapport à la méthode de Dunning. Les modèles de langues gé-

Langues	C & T	Dunning	Giguet	Teahan
Français	47.78 56.40	40.70 52.92	46.90 57.40	71.82 75.85
Anglais	43.57 54.42	47.11 57.68	45.34 54.60	80.71 *82.03*
Finnois	72.11 74.63	69.87 74.78	71.47 72.05	83.33 81.63
Hollandais	87.39 53.47	85.71 61.63	88.23 50.72	*94.96* 75.84
Portugais	24.55 35.08	27.16 39.39	21.63 32.53	**57.04** **64.80**
Espagnol	28.97 38.67	30.67 42.13	28.61 39.01	61.33 65.08
Italien	47.98 55.11	47.85 56.44	49.15 56.76	77.50 79.25
Moyenne	50.34 52.54	49.87 55.00	50.19 51.87	75.24 74.93

TABLE 2 – Précision / F-Mesure pour les quatres méthodes de Cavnar & Trenkle, Dunning, Giguet et Teahan en pourcentages (En **gras**, les résultats les plus faibles, en *italique* les plus imporants).

nérés pour la méthode de Teahan capturent ainsi beaucoup mieux l'essence de la langue. En effet, le classifieur PPM essaye de prédire le caractère suivant une suite de N caractères (dans notre cas quatre, car l'ordre maximum des modèles de Markov est de quatre) : si la prédiction est fausse, il essaye de prédire avec la suite des N-1 caractères précédents.

Les résultats augmentent pour des mots plus longs. En effet, plus de caractères implique plus de données pour l'algorithme, ce qui améliore les probabilités et donc l'identification. Ainsi, sur des mots de 10 caractères, la méthode de Teahan obtient une F-Mesure moyenne de 80.13% et la méthode de Dunning 56.80%.

Néanmoins, lors d'une comparaison plus fine des résultats, on se rend compte que souvent, lorsque le classifieur PPM de Teahan effectue une identification erronée ou indéterminée, le classifieur Bayes de Dunning effectue une identification positive et réciproquement. Nous décidons alors qu'il serait intéressant de combiner les résultats de ces deux classifieurs. L'objectif est d'améliorer les résultats, notamment pour les langues portugaise et espagnole (ainsi que française et italienne) possédant des bases linguistiques communes et ayant les résultats les plus bas.

4 Méthode Hybride PPM / N-gramme

Auparavant, de bons résultats ont été obtenus sur des textes très courts, notamment dans les travaux de (Rehurek and Kolkus, 2009) avec une précision moyenne de 80% environ sur leur échantillon *small* (de deux à cinq mots donc plus que dans nos conditions de test) avec un pool de neuf langues, ou dans (Vatanen et al., 2010) avec une moyenne d'environ 65% d'identification positive sur des textes de test de sept caractères (soit environ deux mots) avec un pool de 281 langues. Les résultats de (Rehurek and Kolkus, 2009) sont toutefois à relativiser car les contraintes imposées sont beaucoup plus fortes que les nôtres, de même (Vatanen et al., 2010) possède un pool de langue beaucoup plus important et a été entrainé sur un corpus d'apprentissage plus petit d'une longueur médiane de onze mille caractères. Nous avons décidé de mettre en place une nouvelle méthode en nous appuyant seulement sur des méthodes simples plus anciennes et qui ont servis de base pour toutes les méthodes qui ont suivi. Comme expliqué précédemment, l'idée de base est de coupler les scores obtenus par les méthodes de Dunning et de Teahan en utilisant des algorithmes de pondération afin de gommer les lacunes de chaque méthode.

La méthode de Dunning possède ainsi trois scores différents, un pour chaque N-gramme (2-grammes, 3-grammes, 4-grammes) :

$$S_{Ngram} = \alpha \cdot S_{2gram} + \beta \cdot S_{3gram} + \gamma \cdot S_{4gram} \quad (10)$$

Le nombre de 4-grammes [3] étant beaucoup plus grand, ceux-ci sont plus discriminants pour l'identification de la langue que les 2-grammes. Après plusieurs simulations, on donne les valeurs 1, 2 et 1.5 pour les paramètres α, γ et β.

Le score de la méthode hybride est simplement déterminé de la façon suivante :

$$S_{Hybrid} = \delta \cdot S_{Ngram} + \lambda \cdot S_{PPM} \quad (11)$$

avec S_{Ngram} possédant un score compris entre 0 et 4.5 pour chaque langue avec $\bar{S}^{L}_{Ngram} = 4.5$ et S_{PPM} possédant un score compris entre 0 et 1 pour chaque langue avec $\bar{S}^{L}_{PPM} = 1$. La méthode PPM obtenant de meilleurs résultats en moyenne (voir figure 2), il semble logique de supposer que λ sera supérieur à δ. On fixe alors $\delta = 1$ pour effectuer la calibration λ (voir figure 4), puis nous déterminons $\delta = 1$ et $\lambda = 14$ pour la suite. En effet, d'après nos essais, nous observons que la précision augmentent jusqu'à $\lambda = 7$ environ alors que la valeur de la F-Mesure atteint son maximum pour $\lambda = 14$. La diminution de la F-Mesure entre 14 et 30 est due à la forte diminution du rappel malgré l'augmentation de la précision. En effet, cette méthode a pour objectif de corriger principalement les indéterminations de la méthode de Teahan. La valeur choisie favorise ainsi au maximum la F-Mesure, ce qui a pour objectif d'optimiser au maximum le rapport entre la précision et le rappel.

Dans la suite, nous présentons les résultats de la méthode hybride utilisant le calcul de score précédent dans les mêmes conditions de test que décrite dans le Tableau 1.

3. Pour un alphabet de 26 lettres (donc sans considérer les accents), le nombre de 1, 2, 3 et 4-gramme est respectivement de 26, 325, 2600, 14950

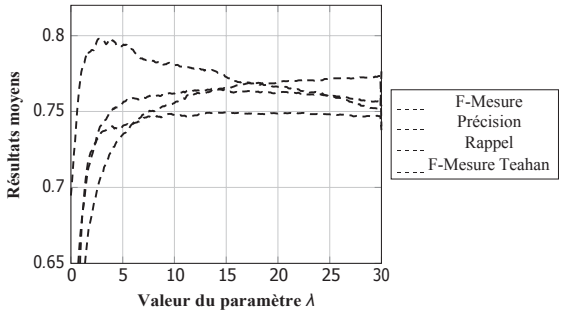


FIGURE 3 – Variation des résultats en fonction de la valeur du paramètre λ

5 Expérimentations et résultats

Les améliorations observées lors des expérimentations sont principalement dues aux corrections des indéterminations de chaque méthode. En effet, si les résultats originaux obtenus par les deux méthodes étaient des indéterminations, en sommant les scores obtenus à l'aide de l'équation 11, on agrandit les différences de score entre les deux meilleures langues, ce qui conduit cette différence à passer en dessous du seuil d'indétermination. Cela entraîne généralement une identification positive, mais aussi une identification négative si les deux méthodes se trompent originellement.

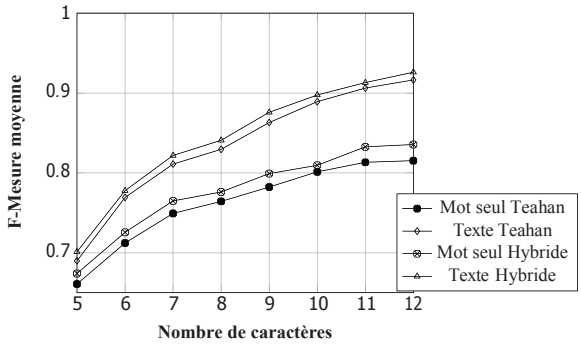


FIGURE 4 – Résultats moyens de la méthode Hybride et de la méthode de Teahan sur des mots seuls et des textes en fonction du nombre de caractères.

Les résultats obtenus sont alors présentés dans les figures 5, 5 et 6. Notre algorithme améliore de 1,57% en moyenne la F-Mesure obtenue par la méthode de Teahan sur des mots seuls tout en améliorant la précision de 0.79%. Ces améliorations s'expliquent notamment par la transformation d'un résultat erroné en résultat indéterminé lorsque les deux méthodes confirment l'indétermination et lors de la transformation d'un résultat indéterminé en identification positive lorsque la combinaison des scores des méthodes enlève l'indétermination. Les résultats sur les langues moins bien identifiées précédemment tels que l'espagnol et le portugais ont subi une amélioration de la précision accompagnée d'une amélioration légère de la F-Mesure. L'objectif voulu est atteint puisque en paramétrant ainsi notre seuil d'indétermination, nous avons mis en avant une identification sans erreur sur des mots seuls afin de pouvoir identifier les mots indéterminés. La nouvelle méthode permet ainsi de mieux différencier les langues proches lors de l'identification de mots seuls, de même que sur des textes contenant jusqu'à 5 mots. Il faudra néanmoins effectuer des tests à plus grande échelle avec un nombre de langues plus grand pour confirmer ces résultats. Comme indiqué sur la figure 5, on voit bien l'amélioration en fonction du nombre de caractères mais aussi entre la méthode de Teahan et la méthode hybride. Sur des mots de 10 caractères, la méthode hybride obtient une F-Mesure moyenne de 80.96% soit une amélioration de 0.82% par rapport à la méthode de Teahan. Les résultats obtenus montrent qu'en combinant simplement les méthodes existantes, il est possible d'obtenir des résultats globaux plus intéressants.

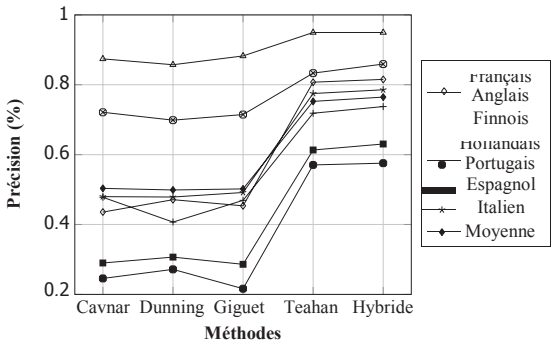


FIGURE 5 – Précision pour la méthode hybride comparé aux méthodes de l'état de l'art.

Afin d'effectuer une comparaison réelle avec les résultats les plus récents concernant l'identification de tweet (Panich, 2015), nous avons tester notre algorithme sur le corpus TweetLID [4] (voir tableau 5) et seulement pour les langues apprises. Ainsi, lors de nos tests, nous avons retiré les tweets meilleur comme observé sur le tableau 5 mais celui-ci est biaisé par le retrait de ces langues, linguistiquement très proche, proche aussi de l'es-

4. http ://komunitatea.elhuyar.eus/tweetlid

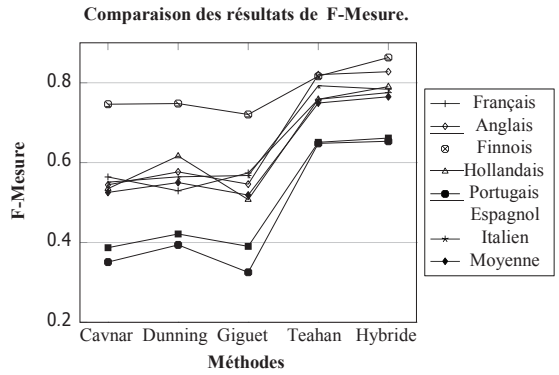


FIGURE 6 – F-Mesure pour la méthode hybride comparé aux méthodes de l'état de l'art.

Méthode	F-Mesure
Hybride	88.83
Improved graph-based N-gram approach	83.63*
N-gram approach with the naive Bayesian classifier	≈82*

TABLE 3 – Comparaison des résultats sur le corpus TweetLID 2014. * indique que le résultat a été obtenu sur le corpus complet.

pagnol. Ainsi, sur textes courts contenant des erreurs typographiques, notre méthode obtient aussi de bons résultats.

6 Conclusion et perspectives

Dans cet article, nous avons présenté nos travaux relatifs à un nouvel algorithme de classification pour le traitement automatique de la langue pour des textes courts et très courts. Tout d'abord, nous avons testé les classifieurs les plus représentatifs et fondateurs dans la littérature (ayant servis de socle pour plusieurs travaux qui les ont succédés) : les résultats obtenus ont montré qu'il était possible d'améliorer les différents taux de détermination/erreur/indétermination en combinant plusieurs méthodes (celles qui donnent les meilleures performances) grâce à des poids de pondération fixés après plusieurs tests de calibration.

En choisissant des corpus libres de droit et disponibles sur Internet, nous avons établi des bases de comparaison fournies. Les résultats obtenus montrent que notre méthode hybride améliore en moyenne la F-Mesure sur l'identification d'un seul mot. Néanmoins, ces travaux restent préliminaires et sujet à amélioration. En effet, nous prévoyons de considérer des algorithmes d'optimisation tel qu'un algorithme génétique afin de diminuer le taux d'indétermination dû aux ressemblances entre langues ayant la même racine (*e.g.* espagnol et portugais). Aussi, notre méthode combine 2 classifieurs (i.e. PPM et Bayes) : il serait intéressant de voir l'influence d'autres classifieurs sur les résultats obtenus comme les classifieurs SVM ou Cosine Similarity (Brown, 2013) ou étudier les méthodes utilisant les graphes (Tromp and Pechenizkiy, 2011). D'autre part, nous allons aussi augmenter le pool de langues à détecter afin d'étudier les variations des résultats, notamment sur des langues proches du finnois ou du hollandais mais aussi l'allemand.

References

Brown, R. D. (2013). Selecting and weighting n-grams to identify 1100 languages. In *International Conference on Text, Speech and Dialogue*, pages 475–483. Springer.

Cavnar, W. B. and Trenkle, J. M. (1994). N- gram-based text categorization. *Ann Arbor MI*, 48113(2) :161–175.

Dunning, T. (1994). Statistical identification of language.

Giguet, E. (1995). Categorization according to language : A step toward combining linguistic knowledge and statistic learning. In *Proceedings of the 4th International Workshop on Parsing Technologies (IWPT-1995), Prague, Czech Republic*. Citeseer.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

McNamee, P. (2005). Language identification : a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3) :94–101.

Padró, M. and Padró, L. (2004). Comparing methods for language identification.

Panich, L. (2015). Comparison of language identification techniques.

Rehurek, R. and Kolkus, M. (2009). Language identification on the web : Extending the dictionary method. In *Computational linguistics and intelligent text processing*, pages 357–368. Springer.

Teahan, W. J. (2000). Text classification and segmentation using minimum cross-entropy. In *Content-Based Multimedia Information Acess-Volume 2*, pages 943–961.

Tromp, E. and Pechenizkiy, M. (2011). Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34.

Vatanen, T., Väyrynen, J. J., and Virpioja, S. (2010). Language identification of short text segments with n-grams models. In *LREC*.

Imminence contrecarrée en russe et en français: explication cognitive des différences d'expression grammaticale

Alexandr Ivanov
77, rue Daguerre
75014 Paris

alexander.sergeevich.ivanov@gmail.com

Résumé/Abstract

Dans cet article nous analysons les moyens grammaticaux dont disposent la langue russe et la langue française pour exprimer l'imminence contrecarrée qui fait partie du champ sémantique plus vaste de l'antirésultatif. La langue française dispose de plusieurs verbes modaux et des expressions à sens modal qui servent à exprimer ce sens tandis que le russe possède beaucoup de moyens lexicaux pour exprimer le fait que l'action s'est arrêtée à la proximité imminente du résultat. Ces particules sont souvent rajoutées aux verbes modaux pour nuancer encore plus le sens de l'imminence contrecarrée. Nous pouvons supposer que cette différence provient de la façon de focaliser l'attention sur les phases différentes du déroulement interne du procès. Dans cet article nous analyserons les différences de l'expression grammaticale de l'imminence contrecarrée en utilisant le modèle de R. Langacker de profiliation (profiling). Nous essayerons de combiner l'explication sémantique basée sur les travaux de V. Plungian, Wilmet, Nedljakov etc, avec le modèle cognitif de R. Langacker. Mots-clés : linguistique cognitive, antirésultatif, proximatif, périphrase verbale, réel, irréel.

1 Introduction

Nombreux sont les travaux qui traitent de la notion du résultat comme catégorie verbale. On peut dire qu'elle est probablement l'une des catégories les mieux étudiées du point de vue de l'aspect et de la typologie. De nombreux travaux ont été publiés sur cette question. Notamment des linguistes et philologues russes qui se concentrent sur la catégorie de l'aspect verbal (qu'ils appellent vid) et surtout sur les énoncés marquant le résultat. Parmi les ouvrages les plus connus dans ce domaine on peut citer ceux de V. P. Nedljakov «Typologies des constructions résultatives » 1983, Coseriu 1976. Parmi les travaux les plus proches à notre étude on peut citer ceux de V. Plungyan l'article « Avant et après le résultat » où il essaie de remplir les « lacunes » liées à l'absence des travaux pointus sur le sujet de l'imminence contrecarrée. Les linguistes français se sont intéressés au sujet du résultatif ce qui donne lieu à des travaux de classification des aspects ainsi qu'au travaux décrivant la constructions exprimant le résultat. Notemment les guillaumiens : G. Moignet, M. Wilmet.

2 Cadre théorique

La notion de l'imminence contrecarrée est étroitement liée à la notion du résultat ou surtout à la notion d'aboutissement. Cette notion est fondamentale pour la compréhension de l'imminence contrecarrée. Dans la langue française l'aboutissement à un résultat suppose un procès dirigé vers un but qui soit le point final. Cela suppose une certaine volonté à atteindre le résultat. La catégorie verbale de résultat suppose que dans la langue il y a des formes lexico- morphologiques au moyen desquelles la catégorie du résultat est véhiculée. Ces formes servent à décrire les situations d'aboutissement qui impliquent la notion de limite. L'aboutissement suppose qu'il y a une limite ou un terme quelconque qui arrête le procès

et au-delà duquel le procès ne continue plus ayant atteint son terme naturel. En même temps le terme d'aboutissement implique que c'est vers ce terme que le procès est dirigé, alors ce terme naturel non seulement sert du point ad-quem du procès au-delà duquel il ne continue plus, mais aussi représente un point vers lequel on dirige une certaine volonté. La notion d'aboutissement ne fait pas partie intégrante de la sémantique verbale. Les verbes peuvent se référer aux situations qui ont un terme naturelle ou non. Les verbes qui se réfèrent aux situations qui ont ce terme sont appelés téliques tandis que les verbes qui se réfèrent aux situation sans ce terme sont atéliques.

Le résultat étant une catégorie aspectuelle est étroitement lié à l'aspect lexical du verbe. Le résultat peut être atteint seulement dans certains groupes de verbes : les verbes téliques c'est-à-dire les procès qui sous la condition du déroulement normal du procès peuvent aboutir, atteindre un résultat quelconque. Il est à noter que le résultatif décrit la fin du déroulement d'un procès ininterrompu. Tout développement d'une situation du monde peut se composer des étapes suivantes : a) l'état initial (ou d'avant le procès), b) le début, c)le développement, d) la fin et e) l'état final (ou d'après le procès).
A partir de ce schéma nous pouvons voir qu'il y a trois points ou le déroulement normal du procès peut être perturbé :

1 Le point final n'est pas atteint, le procès avait été interrompu dans sa phase du déroulement.
2 L'état du résultat avait eu lieu, mais pour une raison quelconque a été annulé.

Pour le premier et le deuxième point il y a plusieurs marqueurs dans les langues du monde et notamment dans la langue française et dans la langue russe. Au cours du développement de ces langues il y a eu des marqueurs différents pour exprimer soit le premier, soit le deuxième point en même temps ou séparément. Le fait que les deux points sont souvent exprimés dans la langue en même temps est à cause du fait que c'est la phase du résultat qui est perturbé (Plungian, 2001, p.51). On peut regrouper ces deux cas dans la même catégorie, celle de l'antirésultatif. Dans le présent article nous nous intéressons uniquement au premier groupe des antirésultatif: la perturbation du procès avant le résultat. On va comparer comment cette catégorie sémantique est véhiculée par les moyens lexico-grammaticales dans la langue française et dans la langue russe en essayant de donner une explication cognitive concernant les différences dans l'expression de l'imminence contrecarrée dans ces deux langues indo-européennes qui appartiennent aux groupes linguistiques différents, mais qui ont le statut des langues mondiales et qui ont joué le rôle important (pour des raisons historiques et sociales différentes) dans le monde.

2.1 Antirésultatif: paramètres

Il est évident pour ceux qui connaissent même un peu le français que la zone sémantique de l'antirésultatif est assez vaste dans cette langue. Il faut noter qu'un point commun pour toutes les catégories est le non aboutissement du processus dans son déroulement normal. Pour différencier les nuances de signification sémantique nous serons obligés de prendre en compte les paramètres suivants que Plungian propose dans l'article « Antirésultatif : avant et après le résultat »:

1 La possibilité de continuer le procès après son arrêt
2 la volonté du sujet d'atteindre le résultat
3 la proximité au résultat (degré d'étroitesse).

Le premier paramètre sert à distinguer les procès qui n'ont pas abouti des procès qui ont été arrêtés au cours de leur déroulement, mais qui peuvent continuer. Les processus décrits ici sont bornés par les limites extérieures qui ne relèvent pas de leur aboutissement naturel. Les procès relevant de ce paramètre ne nous intéressent pas dans le présent article, car il n'expriment pas le non-aboutissement, mais plutôt un arrêt dans une phase de déroulement. Le bornage externe n'est pas un aboutissement mais un arrêt du déroulement du processus.
Les deux autres groupes ont une valeur sémantique assez proche l'un de l'autre et sont souvent regroupés sous le nom de l'imminence contrecarrée qui exprime bien l'aspect imminent (précursif dans la terminologie de M. Wilmet) de l'action qui n'a pas pu se réaliser pour une telle ou telle raison.
Le troisième paramètre distingue les procès qui se sont arrêtés au point très proche du résultat. En français ils sont d'habitude exprimés par l'adverbe « presque ». Le point important qui distingue ce groupe du deuxième est la polysémie de ce groupe de procès. Dans la phrase: « il a presque ouvert la fenêtre » , il peut s'agir du fait que le procès d'ouverture de la fenêtre n'a pas eu lieu, qu'il s'est arrêté juste avant que la fenêtre

soit ouverte ou peut marquer un arrêt dans le processus d'ouverture de la fenêtre qui ne manquera pas d'aboutir dans le futur proche. Il s'agit d'antirésultatif qu'au premier cas. Dans l'exemple suivant on peut bien observer l'adverbe dans le sens d'antirésultatif : Alors elle murmura, presque défaillante. (Guy de Maupassant « Bel Ami »).
Le deuxième paramètre divise les situations en deux groupes principaux: ceux qui peuvent être contrôlés et ceux qui ne peuvent pas. Seules les situations contrôlables nous permettent de diriger notre volonté vers leur aboutissement. En cas d'échec cet effort de volonté est considéré comme une tentative. Selon Plungyan, la signification d'une tentative échouée est très souvent grammaticalisée dans les langues. La langue française ainsi que la langue russe préfèrent recourir aux verbes modaux qui montrent bien la volonté dirigée vers l'aboutissement du processus :

L'autre s'en est rendu compte et a voulu s'éloigner…, mais l'Inconnu … l'a si bien atteint qu'il lui a brisé le crâne. (Renaud de Beaujeu « Le Bel Inconnu »).

Ya khotel molcha proehat mimo ih, no oni menya totchas okroujili/ J'ai voulu passer à côté d'eux, mais ils m'ont de suite entouré (Pouchkine « La fille du capitaine »)[1].

L'autre groupe se rapporte aux procès qui ne peuvent pas être contrôlés et il décrit les procès qui s'arrêtent (soudainement, indépendamment de la volonté du sujet) en proximité imminente du résultat du procès. On va appeler cette catégorie les proximatifs suivant le terme donné par Comrie (Comrie,1976, p.64). Le représentant le plus classique de cette catégorie en français moderne est le verbe faillir au passé : Elle faillit tomber (Zola « Assommoir »). Il faut noter que les proximatifs se trouvent plutôt en périphérie de la zone sémantique d'antirésultatif et très souvent les auxiliaires ou les périphrases qui servent à exprimer non seulement le résultat non atteint mais aussi le fait que le résultat n'est pas voulu ou n'est pas attendu :
ils avaient failli crever de rage (Zola « Assommoir »).

Le troisième groupe implique alors une modalité qui découpe le champ sémantique de l'antirésultatif en deux parties : celle de l'existant et celle de l'inexistant. C'est à cette coupure modale que s'articule l'expression de l'imminence contrecarrée. L'imminence contrecarrée se trouve aux abords de la coupure modale. Deux cas de figure sont possibles : 1) l'effort/la volonté du sujet est dirigée vers un résultat voulu et a été interrompu dans la proximité imminente indépendamment de la volonté du sujet. 2) le résultat non-voulu n'a pas été atteint soit à cause d'un effort du sujet soit non, ce qui nous renvoie à la notion des procès contrôlables et non-contrôlables. La notion des procès contrôlables est empruntée aux linguistes slavisants, mais dans notre étude elle présente une grande importance car c'est sur ce que se trouve la grande différence entre l'expression de l'imminence contrecarrée en russe et en français.

2.2 Imminence contrecarrée du point de vue de la linguistique cognitive

Selon la conception de R. Langacker chaque verbe et expression modaux font référence à une autre activité (Langacker, 1991, p.270). Ainsi les verbes ou les expressions se réfèrent à un autre procès qui sert de landmark (repère pour l'action décrite). L'action du sujet représente une certaine potentialité envers le processus – repère (landmark). Néanmoins, la réalisation de cette potentialité n'est pas actuelle, mais potentielle. Si on fait l'abstraction des exemples individuels et considérons les modaux comme une classe, on pourra observer la différence principale entre les deux groupes des modaux. En ajoutant de la subjectivité dans l'action envisagée, nous pouvons voir que la différence entre les verbes et les expressions modaux se résume à la différence de profiler soit l'action- repère, soit les relations modales. Ainsi les expressions modales françaises et russes se focalisent sur les relations modales selon la conception de R. Langacker, car ils possèdent plusieurs caractéristiques sémantico-lexicales qui permettent de les classer dans ce groupe (notamment la conjugaison et l'infinitif). La différence entre la langue russe et la langue française se trouve non dans la façon de voir la potentialité, mais dans la façon de focaliser l'attention (profiler) telle ou telle partie de la situation.
Cela peut être illustrée par un schéma ci-dessous :

[1]Я хотел молча проехать мимо их; но они меня тотчас окружили (C'est moi qui traduis)

La flèche « a » au-dessous de la ligne de situation symbolise le déroulement du processus interrompu par un événement qui vient s'interposer dans son déroulement naturel opérant la coupure modale au-delà de laquelle se trouve le résultat naturel du processus vers lequel l'énergie est dirigée représenté par la flèche « b ». La flèche « c » représente le degré d'étroitesse, la distance entre le réel et l'irréel.

3 Imminence contrecarrée en russe et en français

Ainsi la langue française moderne a tendance à profiler l'énergie de la situation qui est dirigée vers la fin logique du processus. Dans la phrase comme : « Gervaise qui juste allait acheter un sou d'oignons brûlés … fut prise d'un tremblement et n'osa plus sortir» (E. Zola «Assomoir »). On peut voir que l'énoncé profile l'effort dirigé vers l'accomplissement du processus tandis que le degré d'étroitesse sert de base de cette expression.

La langue russe, contrairement à la langue française exprime rarement l'imminence contrecarrée seulement à l'aide des verbes modaux. Les expressions utilisées sont « chut ne » (à peu que ne , équivalent à « pour poi que » en ancien français) et « chut biilo ne » : Dami akhnuli, nekotoriie daje chut ne upali v obmorok / Les dames ont poussé un cri, certaines ont failli défaillir (Akounine)[2].

La langue russe utilise plus les moyens lexicaux pour exprimer l'imminence contrecarrée en centrant l'attention sur le degré d'étroitesse plutôt que sur la coupure modale. Bien que les deux paramètres soient présents dans l'expression de l'imminence contrecarrée dans les deux langues, elles concentrent leur attention sur l'un des deux. En suivant le modèle épistémologique de R. Langacker on peut dire que les expressions traduisant l'imminence contrecarrée se trouvent dans le champ d'irréel mais avec une proximité de la zone de réalité dans le passé. La différence que l'on peut voir entres les deux langues est leur façon de conceptualiser cette distance. La langue française à plutôt la tendance à marquer l'aspect d'une action non-réalisée en focalisant l'attention sur la flèche d'effort ou de volonté tandis que le russe a plutôt la tendance à marquer le degré d'étroitesse. Les locuteurs du russe focalisent leur attention plus sur le fait que l'action a été interrompue à la proximité imminente de sa fin. La langue russe renforce l'irréalité du processus décrit par l'emploi plus fréquent de la particule de négativité « ne ». Dans une phrase comme : « Rykalov perepolochilsya, chut stoul ne oprokinoul /Rykalov a eu peur, et a failli faire tomber la chaise » (Acounine). [3] Nous pouvons voir qu'on nous présente positivement avec la particule peu (chut) le procès dont le résultat est à la fin négatif (il ne fait pas tomber la chaise). C'est le degré d'étroitesse qui est mis en avant laissant l'énergie du procès en arrière-plan.

Il est à noter que la langue russe se sert des verbes *vouloir* et *croire* pour marquer l'imminence contrecarrée, mais leur emploi est souvent accompagné des adverbes d'intensification (ujé) ou la particule du conditionnel (biilo). Dans les exemples suivants on peut bien voir que les verbes modaux russes présentent beaucoup de ressemblances avec les verbes modaux français. La différence se trouve dans la manière d'expression de l'irréel qui est beaucoup plus accentué en russe qu'en français, dont l'emploi constant des particules du conditionnel lors de l'expression de l'imminence contrecarrée :

Doug hotel biilo otodvinutsya, no reshil, chto ono togo ne stoit. / Doug a voulu faire un mouvement, mais a décidé que cela ne valait pas la peine (Loukianenko).[4]

Ya otkriil komnatu, vkluchil svet i sobralsya biilo uje viigruzit produkti na stol, kogda za spinoi chto-to goulko khlopnulo / J'ai ouvert ma chambre, allumé la lumière et étais près de mettre tous les produits sur la table quand quelque chose a bruyamment éclaté (Loukianenko).[5]

Dans la langue française cette coupure modale est véhiculée par les auxiliaires modaux qui forment le noyau dur de ce champ de la proximité éventuelle au résultat non-atteint (virtuel). Dans la langue russe, les périphrases verbales sont encore moins grammaticalisées et ne présentent pas une entité constante. La langue française possède les moyens d'expression de l'imminence contrecarrée à l'aide des verbes modaux :

[2]Дамы ахнули, некоторые даже чуть не упали в обморок (C'est moi qui traduis)

[3]Рыкалов переполошился, чуть стул не опрокинул (C'est moi qui traduis) .

[4]Дуг хотел было отодвинуться, но решил, что оно того не стоит (C'est moi qui traduis) .

[5]Я открыл свою комнату, включил свет и собрался было уже выгрузить продукты на стол, когда за спиной что-то гулко хлопнуло (C'est moi qui traduis) .

vouloir, croire. Dans la langue française les verbes modaux sont employés pour marquer la coupure modale entre le réel et l'irréel qui s'opère lors de la construction de l'énoncé. Dans ce sens il montre des similarités avec les périphrases verbales.

4 Conclusion

En conclusion, on peut voir que la langue française et la langue russe possèdent bien des expressions qui désignent l'imminence contrecarrée, mais la façon dont les langues expriment cette notion est différente. La langue française se focalise sur l'intention, elle possède beaucoup de marqueurs spéciaux pour l'antirésultatif tandis que la langue russe possède plusieurs marqueurs lexicaux qui permettent de nuancer le degré d'étroitesse c'est-à-dire le fait que l'action s'est arrêtée à une proximité imminente du résultat. On peut voir que le français a tendance a plus grammaticaliser les expressions traduisant l'imminence contrecarrée que le russe qui emploie plus de moyens lexicaux pour exprimer cette notion.

Références/References

Comrie, B. (2001). *Aspect*. Cambridge [u.a.]: Cambridge Univ. Pr.

Coseriu, E. & Bertsch, H. (1976). *Das romanische Verbalsystem*. Tübingen: TBL.

Garey, H. (1957). Verbal Aspect in French. *Language*, *33*(2), 91-110. doi:10.2307/410722

Guillaume, G. (1965). *Temps et verbe : théorie des aspects, es modes et des temps* (2nd ed.). Paris: Librairie Honoré Champion.

Plunguian, V. (2001). Антирезультатив : до и после результата. *Исследования По Теории Грамматики*, (1), 50-58.

Langacker, R. (1991). *Foundations of Cognitive Grammar* (1st ed.). Stanford: Stanford University Press.

Langacker, R. (1987). *Foundations of cognitive grammar* (1st ed.). Stanford, Calif.: Stanford University Press.

Langacker, R. (2008). *Cognitive grammar* (1st ed.). Oxford: Oxford University Press.

Maslov, U. (2004). *Очерки по аспектологии* (2nd ed., pp. 18-302). Москва: Языки славянской культуры.

Milliaressi, T. (2015). *Aspects et temporalité* (1st ed.). Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Nedljakov, V. (1983). *Типология результативных конструкций (результатив, статив, пассив, перфект)* (1st ed.). Ленинград: Наука.

Wilmet, M. (2010). *Grammaire critique du français* (1st ed.). Bruxelles: De Boeck Duculot.

More experiments with the Tag Thunder concept

Elena Manishina,
IRIT (UT3)
18 Route de Narbonne
F-31062 Toulouse
elena.manishina@irit.fr

Fabrice Maurel, Jean-Marc Lecarpentier, Stéphane Ferrari
Normandie Univ, UNICAEN,
ENSICAEN, CNRS, GREYC,
14000 Caen
firstname.lastname@unicaen.fr

Abstract

Tag cloud is a resume of a web page content which groups the key terms presented using typographic effects and reflecting their relevance for a given page. A tag thunder is an audio version of a tag cloud. In tag thunders the relevance of a given key term is translated into specific speech effects and its position on the page is reflected in the position of the corresponding sound on a 2D stereo space. Tag thunders serve to provide speed reading techniques in non-visual web browsing environments and allow visually impaired users to get a quick glimpse of the web page content without the need to read through the page. The first evaluation results of our implementation of the tag thunder concept demonstrated its potential and viability as a non-visual alternative to visual speed reading techniques. In this paper, we present the experimental results of the second stage of the evaluation campaign where we assess the quality of our vocalization strategies and its impact on the content perception and understanding by the users.

Keywords : non-visual web navigation, human-computer interaction, text-to-speech synthesis

1 Introduction

When it comes to quick browsing of the web content, such document properties as layout, logical structure and typographic effects play an important role in the perception process. However, these properties are usually not rendered in non-visual browsing setup. Figure 1 illustrates how a web page is rendered in visual and non-visual setups. Most of the existing solutions ([Borodin et al., 2010; Ahmed et al., 2012]) however, do not fully provide the capabilities of the visual browsing environment. Our solution, which we call Tag Thunders (TT), provides skimming (quick reading) techniques for non-visual browsing. A tag thunder is the vocal equivalent of the tag cloud concept. Unlike tag clouds, where key terms are presented using typographic effects which reflect their relevance and number of occurrences, tag thunders use specific speech effects and 2D stereo spatialisation to represent the relevance of a given key term and its position on the page.

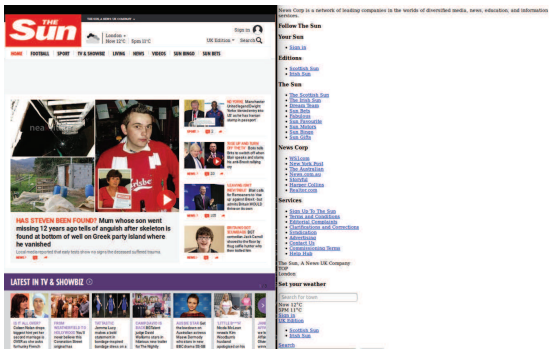


FIGURE 1: Perception of the same web page in visual and non-visual environments.

Tag thunders provide a 'skim' of web pages, thus giving users the general information about the web page content and layout, and allowing for further navigation within the page.

The tag thunder generation process unfolds as follows : first, given an input url, cleaning and visual information extraction are performed. Second, the web page is segmented into a given number of zones (5 in our current setup) ; as a rule the resulting segmentation reflects the logical structure of the page : the menu, the main content, which might in turn be split into two or more zones, the footer, the side menus, etc.

Then, for each zone, we extract key terms which would represent the zone in the tag thunder. The keyword extraction relies on several selection cri-

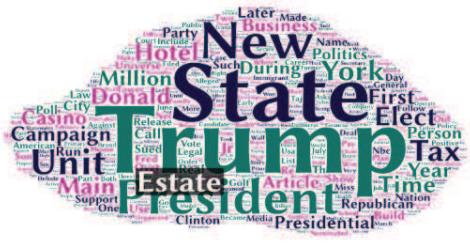


FIGURE 2: Example of a tag cloud.

teria, specifically, one of the major ones is the visual properties of the zone and its content. Extracting and vocalization of visually glaring page elements provide the maximum resemblance to visual web page skimming. Finally, the extracted key terms are simultaneously placed on an audio track in order to reflect their positions and visual properties.

To assess the performance of our TT implementation we organized a 3-stage user evaluation campaign. At each stage we evaluate a specific system property, as well as the overall user experience and satisfaction.

In this article we present the results of the second stage of the evaluation campaign (Section 4). The results and discussion of the first stage can be found in [Lecarpentier et al., 2016].

This paper is structured as follows. Section 2 briefly presents our implementation of the tag thunder concept ; for the full description of the TT generation pipeline see [Lecarpentier et al., 2016] and [Manishina et al., 2016]. Section 3 introduces the three-step evaluation campaign which aims at assessing the performance of our tag thunder implementation from different perspectives and angles and providing a thorough understanding of users needs. In Section 4 we discuss the results of the second evaluation stage. We conclude this paper with a discussion and some directions for our future work in section 5.

2 Architecture

In this section we briefly present our implementation of the Tag Thunder concept. It contains three modules : web page segmentation, key term extraction and key term vocalization within the concurrent speech paradigm.

For webpage segmentation we chose an unsupervised learning algorithm (K-means++) [Arthur and Vassilvitskii, 2007] [1]. The algorithm groups visible HTML elements into 5 zones based on their distance.

The next step is to extract key terms from each zone which would represent this zone in the tag thunder. In our current implementation, key terms are n-grams of different lengths with a maximum order of 6 and each zone is represented by one key term only.

Finally we generate the audio signal for each key term taking into account its zone properties. To vocalize the terms we use the Kali TTS [Morel and Lacheret-Dujour, 2001] tool, developed at the University of Caen Normandie by the CRISCO laboratory. Voices are equally distributed on the 2D stereo space according to the zone's centroid coordinates.

3 Evaluation campaign

In this Section, we briefly present the experimental setting and the results of the first evaluation stage. The objective of this experiment is to measure the similarity between web pages and corresponding tag thunders as well as the capacity of tag thunders to give a comprehensive overview of the page.

The experiment unfolds as follows : a user sees a tag cloud followed by a web page, 15 seconds each. The page may or may not be the corresponding web page. The user is then asked whether the tag cloud corresponded to the displayed page. Another user is presented with the same set of keywords, but in the form of a tag thunder instead of a tag cloud ; he/she is then asked to answer the same question.

The results of the first evaluation stage showed that the participants were able to correctly establish the correspondence between a tag cloud and a web page in 87% and the correspondence between a tag thunder and a web page in 84% of cases. The results of TTs being comparable in the overall accuracy with the results of the TCs, we were able to conclude that tag thunders manage to reproduce the visual information representation in the audio format with sufficient precision.

4 Second evaluation stage

In this section we present the results of the second stage of the evaluation campaign. Our goal

1. http://scikit-learn.org/stable/modules/clustering.html\#k-means

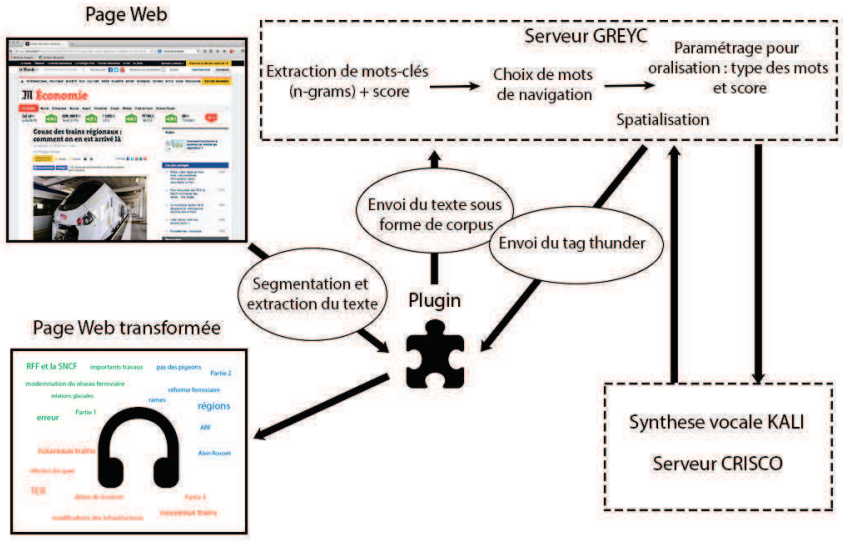


FIGURE 3: Software architecture

here is to evaluate the system's capacity to provide comprehensible web page gists. Specifically we want to evaluate the quality of the generated tag thunders and to analyze the capacity of users to perceive and understand them.

4.1 Experimental setting

The experiment unfolds as follows. Participants hear a tag thunder generated for a particular web page for 15 seconds. They are then asked to write down the words/phrases that they retained. The experiment modalities were as follows :

— 18 sighted participants
— 12 web pages from various web sites were used to generate tag thunders for each page ;

Each page has been tested by at least three users. The participants were given a set of stereo headphones in order to ensure the proper placement of sounds on the stereo space and the proper entry point (left/right ear).

4.2 Results

Table 1 summarizes the results in terms of precision and recall. In our case precision is the proportion of correctly identified key terms among those provided by users and recall is the number (proportion) of terms identified by users out of all terms extracted by the system from a given page. Figure 4 gives an example of the keywords extracted by our system from one of the test web pages and vocalized in a tag thunder.

As we can see from Table 1 there is no direct correlation between the average length of key

handicap visuel
un film pour mieux comprendre
faire un don
lecture sonore
principales maladies de la vue

FIGURE 4: A list of keywords extracted from one of the test web pages

phrases, precision and recall. So for most users there seems to be no difference in perception of the keyword "portfolio" and "Les symptômes de la dépression" : both are correctly identified.

The major problem turned out to be uncommon words like names and terms specific for a given web site ; also slang words and generally unusual (not widely used) key phrases are not identified. For example terms like "weblogs", "guerreiro", "top des qr", etc. have not been recognized by most users. On the contrary, fixed phrases and expressions, as well as commonly used terms are generally correctly identified by most users.

The overall precision of 0.83 may indicate a relatively good quality of the vocalization and the output TTs. A rather low overall recall may be interpreted in three different ways :

— the presence of unknown elements, like names, slang words, etc. in the key phrases extracted from the web page impedes the perception process ;
— perception specificities of each particular user influence the perception process ;

PageID	1	2	3	4	5	6	7	8	9	10	11	12
AKwL (words)	3.6	1.8	1.4	3.4	1.4	2.6	2.8	2.6	2.2	3.4	4.4	1.8
Precision	0.88	0.88	1.0	0.91	0.85	0.83	0.96	0.88	0.75	0.62	0.5	0.89
Recall	0.46	0.63	0.73	0.6	0.5	0.46	0.77	0.7	0.57	0.57	0.23	0.53

TABLE 1: Average keyword length (AKwL), precision and recall per page

Precision	Recall	F-score
0.83	0.56	0.67

TABLE 2: Overall precision and recall

— a high number of key terms vocalized at once (5 in our setup) may be hard to identify ;
— the quality of vocalization (voice settings, like pitch, pace, prosodic patterns, are not well suited for a given setup or a given user) ;
— other issues (to be analyzed with a closer examination) ;

A more precise explanation requires further experiments and a deeper analysis of the user output.

5 Conclusion

In this article, we presented the results of the second stage of the evaluation campaign that we organized in order to test our implementation of the tag thunder concept. These results show that the participants were able to correctly identify most of the vocalized key terms. The results also demonstrate a sufficient quality of the generated audio tracks leaving at the same time some issues to be examined and addressed in the future. The next step is the evaluation of our TT generator with visually impaired participants and using their feedback to direct our future work.

6 Acknowledgments

This research work was funded by the 'Region Normandie' with the CPER NUMNIE project.

7 Website

Tag thunder generator : `https://tagthunder.greyc.fr/demo/`
Experiment (French version) : `https://tagthunder.greyc.fr/demotest`

References

Ahmed, F., Borodin, Y., Soviak, A., Islam, M., Ramakrishnan, I., and Hedgpeth, T. (2012). Accessible skimming : Faster screen reading of web pages. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 367–378.

Arthur, D. and Vassilvitskii, S. (2007). k-means++ : The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Borodin, Y., Bigham, J. P., Dausch, G., and Ramakrishnan, I. (2010). More than meets the eye : A survey of screen-reader browsing strategies. In *Proceedings of the International Cross Disciplinary Conference on Web Accessibility (W4A)*, pages 1–10.

Lecarpentier, J.-M., Manishina, E., Maurel, F., Ferrari, S., Giguet, E., Dias, G., and Busson, M. (2016). Tag thunder : Web page skimming in non visual environment using concurrent speech. In *Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 1–8.

Manishina, E., Lecarpentier, J.-M., Maurel, F., Ferrari, S., and Busson, M. (2016). Tag thunder : Towards non-visual web page skimming. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.

Morel, M. and Lacheret-Dujour, A. (2001). Kali, synthèse vocale à partir du texte : de la conception à la mise en oeuvre. *Traitement Automatique des Langues 42*, pages 193–221.

PROGRAMME du Jeudi 18 mai 2017

Maison de la recherche – Amphithéâtre F417

8:30	**9:00**	Accueil des participants / *Registration*
9:00	**9:30**	Discours de bienvenue / *Welcoming speech*
9:30	**10:30**	**Conférencière invitée / *Conference by guest speaker*** **Audrey Bürki** (Université de Postdam, Allemagne) Interface oral/écrit, ou le rôle du langage écrit dans la production et la reconnaissance des mots
10:30	**11:00**	Pause-café / *Coffee break*
11:00	**12:00**	**Session communications orales 1 / *Oral session 1*** **Redouane Bougchiche** (Université Paris 4-Paris Sorbonne, France) Langue, locuteur et analogie dans l'acquisition-apprentissage linguistique **Wenjia Cai** (University of Edinburgh, Ecosse) First language attrition at two interfaces: binding interpretations of *ziji* "self" by Chinese-English bilinguals
12:00	**14:00**	Pause déjeuner / *Lunch break* (maison de la recherche salle E412)
14:00	**16:00**	**Session communication 2 / *Oral session 2*** **Aleksandra Miletic** (Université Toulouse 2 Jean Jaurès, France) Building a morphosyntactic lexicon for Serbian using Wiktionary **Olga Kataeva** (Institut Catholique de Toulouse, France) et **Elena Manishina** (Université Toulouse 3 Paul Sabatier, France) Compass : a parallel French-Russian corpus enriched with morpho-syntactic annotation **Benoît Coiffet** (Université Toulouse 2 Jean Jaurès, France) « Cuisinez chic » : les emplois adverbiaux de l'adjectif en français
16:00	**16:30**	Pause-café / *Coffee break*
16:30	**17:30**	**Session posters 1 / *Poster session 1*** **Reham Marzouk** et **Seham El Kareh** (Alexandria University, Egypte) Morphological ambiguities in Egyptian Arabic Dialect Used in Social Media **Chieko Kawai** (Université de Poitiers, France) Le développement de l'organisation syntaxique et discursive en français L2 dans les productions orales des apprenants japonais : débutants aux avancés **Carolina Nogueira-François** (Université Lille 3, France) La langue maternelle et les langues non maternelles connues comme recours pour la communication en portugais. Une étude de cas. **Divna Petkovic** (Université de Belgrade, Serbie) et **Victor Rabiet** (Université Paris Est, France) L'alternance modale après les constructions impersonnelles *sembler que* – étude préliminaire statistique à une approche TAL **Camille Létang** (Université d'Orléans, France) Paramètres prosodiques et ratificationnels au sein des séquences contributionnelles et modélisation de l'interface sémantique/pragmatique
17:30	**17:45**	Clôture de la journée / *Closing speech*
	20:00	Dîner de gala / *Gala dinner*

PROGRAMME du Vendredi 19 mai 2017

Maison de la recherche – Amphithéâtre F417

9:00	**9:30**	Accueil des participants / *Registration*
9:30	**10:30**	**Conférencière invitée / *Conference by guest speaker*** **Marie Lalier** (Basque Center on Cognition Brain and Language, San Sebastian, Espagne) Développement de la lecture et bilinguisme précoce
10:30	**11:00**	Pause-café / *Coffee break*
11:00	**12:00**	**Session communications orales 3 / *Oral session 3*** **Veronica Garcia-Castro** (University of York, Angleterre/University of Costa Rica, Costa Rica) Prediction of Upcoming Words and Individual Differences in L2 Sentence Processing : an Eye-tracking Study **Stéphane Duchatelez** (Université de Toulon, France) L'interface organisation linguistique/organisation poétique à la lumière de la théorie des actes de langage
12:00	**14:00**	Pause déjeuner / *Lunch break* (maison de la recherche salle E412)
14:00	**16:00**	**Session communication 4 / *Oral session 4*** **Nataly Jahchan** (Université Toulouse 2 Jean Jaurès, France) The Importance of Using Psycholinguistic tools for CNL Evaluations **Joro Ny Aina Ranaivoarison** (Université d'Antananarivo, Madagascar/Université Paris-Est Marne-la-vallée, France) Dictionnaire électronique (DE) des noms simples issus de verbes. Les noms issus des alternances *mp-* ou *f-* **Hélène Flamein** (Université d'Orléans, France) Annotation d'éléments spatialisés dans l'oral transcrit
16:00	**16:30**	Pause-café / *Coffee break*
16:30	**17:30**	**Session posters 2 / *Poster session 2*** **Clément Thibert** (Université et ENS de Lyon, France) De certains usages dans la twittosphère : contribution à une sociolinguistique computationnelle **Valentin Nyzam** (Université Paris 8-Vincennes-Saint-Denis, France) et **Mohamed Slim Ben Mahmoud** (Ecole Nationale de l'Aviation Civile, Toulouse, France) Méthode hybride pour l'identification automatique de la langue sur textes courts et très courts **Alexandr Ivanov** (Université Paris 4-Paris Sorbonne, France) Imminence contrecarrée en russe et en français : explication cognitive des différences d'expression grammaticale **Elena Manishina** (Université Toulouse 3 Paul Sabatier, France), **Fabrice Maurel**, **Jean-Marc Lecarpentier**, et **Stéphane Ferrari** (Université de Normandie-Caen, France) More experiments with the Tag Thunder concept
17:30		Remise des prix (meilleure communication orale et meilleur poster) / A*ward for the best oral communication and best poster* Discours de clôture / *Closing speech*