# 24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)

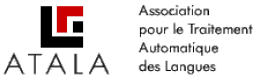
**Orléans, France – 26-30 juin 2017**

https://taln2017.cnrs.fr

# 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)

**Hélène Flamein, Yannick Parmentier (Eds.)**

**Sponsors :**


ATALA
Association
pour le Traitement
Automatique
des Langues

Ortolang

cnrs
dépasser les frontières

eptica
MEANINGFUL CONVERSATIONS

acatus
informatique

aktan

proxem

inbenta

Orléans
Mairie

ORLÉANS
MÉTROPOLE

LOIRE&ORLEANS
ECO

FRENCH TECH
LOIRE
VALLEY

UNIVERSITE D'ORLEANS

Laboratoire
Ligérien de
Linguistique

LABORATOIRE
D'INFORMATIQUE
LIFO
FONDAMENTALE
D'ORLEANS

LI
Laboratoire d'Informatique
EA 2101

Liberté · Égalité · Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE LA CULTURE

# Préface

Vous tenez entre vos mains (ou au bout de votre écran) les actes des 19es Rencontres jeunes Chercheurs en Informatique pour TAL (RECITAL 2017).

Cette année, 17 articles ont été soumis à RECITAL (contre 11 l'année passée), ce qui est encourageant pour la communité du TAL. Ces 17 articles ont tous été relus par 3 membres du Comité de Relecture. Les 5 articles ayant reçu les meilleures évaluations ont été retenus pour présentation orale (soit un taux de sélection pour présentation orale de 29%), et 8 autres articles ont été retenus pour une présentation sous forme de poster (taux de sélection global de 76%).

Nous avons pu ainsi donner l'opportunité à 13 jeunes chercheuses et chercheurs de présenter leurs travaux à la communauté.

Indépendemment du nombre d'articles retenus, RECITAL a à nouveau été l'occasion pour de jeunes chercheuses et chercheurs, de se former à l'écriture d'articles scientifiques, de faire connaître leurs idées et travaux en cours et d'obtenir des avis de la communauté scientifique, autant lors du processus de relecture que, pour les articles acceptés, lors de leur présentation pendant la conférence.

Nous tenons à remercier les nombreux membres du Comité de Relecture pour leur travail sérieux et leurs retours constructifs. Nous espérons que ceux-ci ont pu permettre à tous les auteur.e.s de soumission d'avancer dans leurs travaux.

Merci également à toute l'équipe d'organisation de TALN 2017 (au sens le plus large possible) et aux sponsors pour leur aide dans la réalisation de RECITAL 2017.

Vous souhaitant bonne lecture,

Hélène Flamein, Yannick Parmentier, Co-présidents de RECITAL 2017

# Comités

## Présidente du comité d'organisation de TALN 2017

Iris Eshkol-Taravella (LLL-UMR 7270, Université d'Orléans)

## Vice-président du comité d'organisation de TALN 2017

Jean-Yves Antoine (LI-EA 6300, Université de Tours)

## Président du comité d'organisation de RECITAL 2017

Yannick Parmentier (LIFO-EA 4022, Université d'Orléans)

## Vice-présidente du comité d'organisation de RECITAL 2017

Hélène Flamein (LLL-UMR 7270, Université d'Orléans)

## Présidente du salon d'innovation de TALN 2017

Gabrielle Bosshard (Aktan, Orléans)

## Vice-présidente du salon d'innovation de TALN 2017

Sandra Cestic (LLL-UMR 7270, Acatus Informatique, Orléans)

## Comité d'Organisation

Ahmed Abid (LI, Université de Tours)
Catherine Aléonard (LLL, Université d'Orléans)
Flora Badin (LLL, Université d'Orléans)
Gabriel Bergounioux (LLL, Université d'Orléans)
Sylvie Billot (LIFO, Université d'Orléans)
Marwa Boulakbech (LI, Université de Tours)
Caroline Cance (LLL, Université d'Orléans)
Guillaume Cleuziou (LIFO, Université d'Orléans)
Hyun Jung Kang (LLL, Université d'Orléans)

Anaïs Lefeuvre-Halftermeyer (LIFO, Université d'Orléans)
Denis Maurel(LI, Université de Tours)
Agata Savary (LI, Université de Tours)
Emmanuel Schang (LLL, Université d'Orléans)

## Étudiants du Master Linguistique, spécialité LASTIC de l'Université d'Orléans

Hélène Couderc
Laetitia Delay
Siqi Fan
Lorraine Gaspard
Mélanie Lefeuvre
Sara Masaud
Stéphanie Nogueira
Camille Pertin
Cathy Querineau
Céline Vaschalde
Jidong Xie

## Comité de Relecture de RÉCITAL 2017

Maxime Amblard (Université de Lorraine)
Jean-Yves Antoine (Université François Rabelais de Tours, LI)
Patrice Bellot (Aix-Marseille Université, CNRS, LSIS)
Florian Boudin (Université de Nantes)
Annelies Braffort (LIMSI-CNRS)
Sandra Bringay (LIRMM)
Davide Buscaldi (LIPN, Université Paris 13, Sorbonne Paris Cité)
Marie Candito (Universtité Paris 7, INRIA)
Vincent Claveau (IRISA)
Marcel Cori (MoDyCo, CNRS)
Gaël De-Chalendar (CEA LIST)
Antoine Doucet (University of La Rochelle)
Benjamin Duthil (L3I)
Maud Ehrmann (EPFL, DHLAB)
Karën Fort (Université Paris-Sorbonne)
Thomas François (Université Catholique de Louvain)
Jérôme Goulian (LIG-Getalp)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre, CNRS)
Olivier Kraif (Université Stendhal Grenoble 3)
Mathieu Lafourcade (LIRMM)
Philippe Langlais (Université de Montréal)
Charlotte Lecluze

Benjamin Lecouteux (Université de Grenoble Alpes)
Anaïs Lefeuvre-Halftermeyer (LIFO, Université d'Orléans)
Gaël Lejeune (Université de Nantes)
Laurence Longo (Université de Strasbourg, LiLPa)
Pierre Magistry (Academia Sinica & National Taiwan University)
Jean-Marc Marty (Proxem)
Yann Mathet (GREYC, Université de Caen)
Damien Nouvel (INaLCO)
Alexander Panchenko (Université Catholique de Louvain)
Gaël Patin (XiKO)
Simon Petitjean (Heinrich Heine Universität Düsseldorf)
Jean-Philippe Prost (LIRMM, Université Montpellier 2)
Christian Raymond (UEB/INRIA/IRISA/INSA)
Charlotte Roze (CommunicoTool)
Yves Scherrer (Université Paris 7 Diderot)
Didier Schwab (Université de Grenoble Alpes)
Jérémie Tafforeau (LIF, AMU, CNRS)
Isabelle Tellier (Lattice)
Nadi Tomeh (Université Paris 13, Sorbonne Paris Cité)
Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon, UAPV)
Yannick Toussaint (LORIA, Université de Lorraine)
Mathieu Valette (ERTIM, INALCO, Paris)
Pierre Zweigenbaum (LIMSI, CNRS)

# Table des matières

## Session « oral »

Machine Translation of Speech-Like Texts : Strategies for the Inclusion of Context
*Rachel Bawden* . . . . . 1

Construction de lexiques pour l'extraction des mentions de maladies dans les forums de santé
*Elise Bigeard* . . . . . 15

Création automatique d'une grammaire syntaxico-sémantique
*Emilie Colin* . . . . . 28

Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique
*Yoann Dupont* . . . . . 42

Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires
*Claire Wolfarth* . . . . . 56

## Session « poster »

Générer une grammaire d'arbres adjoints pour l'arabe à partir d'une méta-grammaire
*Cherifa Ben Khelil* . . . . . 70

Déterminants et quantificateurs généralisés dynamiques
*Clément Beysson* . . . . . 81

Détection de l'incertitude et de la négation : un état de l'art
*Clément Dalloux* . . . . . 94

Normalisation de termes complexes par sémantique distributionnelle guidée par une ontologie
*Arnaud Ferré* . . . . . 108

Annotation automatique des lieux dans l'oral spontané transcrit
*Hélène Flamein* ..... 121

Vers détection automatique des affirmations inappropriées dans les articles scientifiques
*Anna Koroleva* ..... 135

Finding Missing Categories in Incomplete Utterances
*Mehdi Mirzapour* ..... 149

Expressions polylexicales verbales : étude de la variabilité en corpus
*Caroline Pasquer* ..... 161

# Machine Translation of Speech-Like Texts: Strategies for the Inclusion of Context

Rachel Bawden
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
rachel.bawden@limsi.fr

ABSTRACT

Whilst the focus of Machine Translation (MT) has for a long time been the translation of planned, written texts, more and more research is being dedicated to translating speech-like texts (informal or spontaneous discourse or dialogue). To achieve high quality and natural translation of speech-like texts, the integration of context is needed, whether it is extra-linguistic (speaker identity, the interaction between speaker and interlocutor) or linguistic (coreference and stylistic phenomena linked to the spontaneous and informal nature of the texts). However, the integration of contextual information in MT systems remains limited in most current systems. In this paper, we present and critique three experiments for the integration of context into a MT system, each focusing on a different type of context and exploiting a different method: adaptation to speaker gender, cross-lingual pronoun prediction and the generation of tag questions from French into English.

RÉSUMÉ

**Traduction automatique de l'« oral-écrit » : Stratégies pour l'intégration du contexte**

Bien que la Traduction Automatique (TA) se soit concentrée jusqu'à présent sur la traduction de textes écrits et édités, de plus en plus de travaux sont consacrés à la traduction de textes informels et spontanés (discours et dialogues). Pour traduire de tels textes relevant de l'« oral-écrit », il devient indispensable de prendre en compte des informations contextuelles, qu'elles soient de nature extra-linguistique (identité du locuteur, interaction entre le locuteur et l'interlocuteur) ou linguistique (coréférence et phénomènes stylistiques propres à la parole). Or l'intégration d'informations contextuelles dans les systèmes de TA reste limitée dans la plupart des systèmes actuels. Dans cet article, nous présentons et analysons trois expériences d'intégration du contexte dans un système de TA mettant en jeu des formes de contexte et donc des méthodologies différentes: l'adaptation au genre du locuteur, la traduction de pronoms et la génération de « tag questions » anglaises à partir du français.

MOTS-CLÉS : traduction automatique, contexte, parole, genre, pronoms, tag questions.

KEYWORDS: machine translation, context, speech-like texts, gender, pronouns, tag questions.

## 1 Introduction

Speech-like texts (social networking, speech transcriptions, subtitles and other informal written exchanges)[1] pose new challenges for Natural Language Processing (NLP) and in particular Machine

[1]Our focus is on the speech-like nature of texts, rather than the medium of communication (written versus oral). The processing of oral discourse is an important research topic, but is not the study of this paper. We therefore choose to refer to the genre of informal, spontaneous productions as being speech-like, whether they are transcriptions of oral discourse or were

Translation (MT), when compared to the genres of parliamentary discourse and journalistic texts, which have been the main focus of text processing applications to date. As the need for high quality MT of this more informal genre increases, more and more MT research is being dedicated to exploring the difficulties speech-like texts present and how we might hope to overcome them (Hardmeier, 2012; Guillou, 2016).

Speech-like texts are fundamentally different from planned written texts for a number of different reasons. One of these is their highly contextualised nature. Whilst it is true that all texts are related to a certain genre and context of production, the extra-linguistic context of a speech-like production is very often essential to understanding its linguistic content and therefore to producing accurate and coherent translations. Unlike journalistic texts, which are addressed to a heterogenous audience, absent at the moment of the text's production, in discourse, the speaker and the interaction with the listener can greatly influence the style (in terms of formality and politeness), vocabulary choices and even the grammatical agreement (in languages that have grammatical gender agreement for example) of the texts. Spontaneity of production allows speakers to continuously adapt to each other's reactions, making the aforementioned contextual aspects even more important. There is also the added possibility that speakers align to each other's linguistic choices, make reference to entities mentioned in the other speaker's speech turn and rely more on implicit common ground to root their speech, potentially resulting in more ambiguity for automatic processing.

Traditional MT methods suffer from a considerable flaw when it comes to dealing with context, this being that sentences are processed independently of each other, and translation choices are often made based on very local context, which is especially true of phrase-based statistical MT (SMT) techniques. Whilst attentional Neural MT (NMT) approaches can partially alleviate this problem, the degree to which we can control which context is used in translation is limited. As for extra-linguistic information, it must first be identified and then integrated into the translation process, through pre-processing, during decoding or as a post-edition step.

We illustrate the different strategies that can be used to integrate contextual information into the MT process by presenting a discussion orientated around three separate experiments for the language pair English-French, each illustrating the integration of a different type of contextual information. The scope and nature of context are very different in each case, presenting different challenges to MT. The experiments described illustrate the nature of the different problems faced, as well as provided a basis for the critique of the methods used, highlighting the fact that the problem is far from being solved.

The first experiment is an illustration of the use of domain adaptation, a very simple technique used to take into account a very coarse-grained and static type of information, that of speaker gender, which applies on the sentence-level (Section 2). The second experiment deals with a finer notion of context, necessitating a more sophisticated modelling of an utterance's linguistic context, in a task to predict the French translation of the English subject pronouns *it* and *they*, which, unlike their equivalent French pronouns, are unmarked for gender (Section 3). In this case, a very specific contextual element (the antecedent) determines the translation of a very specific textual element (the pronoun). Finally, in Section 4, we present our third experiment, a post-editing task to improve the translation of a particular stylistic phenomenon common in spoken English, the English tag question, whose usage is largely determined by speaker attitude and the interaction between speaker and listener. Whilst the phenomenon affected by context (the presence and form of the English tag question) is clearly identified, determining which context can be useful is much more complex.

never produced orally, as in the case of social network posts and discussions.

# 2 Integrating speaker gender through domain adaptation

Speaker identity (as defined by demographic factors such as age, gender and social background, as well as other aspects such as personality and mood) greatly influences the way in which speakers communicate, in terms of lexical choices, syntactic structure, politeness strategies, etc. This variation has long been studied in the sociolinguistics literature, in particular with respect to the impact of gender on communication style (Lakoff, 1975; Coates, 1986). Whilst speaker adaptation has long been a part of speech recognition systems (Kuhn *et al.*, 1998), only recently has research in NLP turned towards adapting systems to take into account this extra-linguistic information (cf. works by Volkova *et al.* (2013) for sentiment analysis and Hovy (2015) for three separate NLP classification tasks). Speaker identity is just as important in MT; van der Wees *et al.* (2016) show how MT performance fluctuates with respect to individual speaker identity, gender and register in five different language pairs for film and TV subtitles, and Mirkin *et al.* (2015) find that MT translated output does not conserve speaker personality and gender traits as well as human translations do.

One of the simplest methods for integrating coarse-grained, sentence-level information such as aspects of speaker identity is to perform domain adaptation. *Domain* in this context refers to a broader concept than the genre of text and can include any sentence-level aspect representing the class of sentence, such as sentence type, speaker gender and formality. Domain adaptation is a well-known approach in MT and is used almost systematically. In its simplest form, it implies simply choosing an in-domain development set to tune the system (cf. Pecina *et al.*, 2012 for a study of the effectiveness of in-domain tuning for SMT systems when dealing with a low-resource domain). But its implementation can be more complex, for example through similarity-based data selection techniques for both training and tuning data (Axelrod *et al.*, 2011) or by weighting different components of the MT system that have been trained on class-partitioned data (Foster & Kuhn, 2007; Finch *et al.*, 2009).

## 2.1 Experiments and results

Our preliminary experiments into the use of domain adaptation to take into account the gender of the speaker (Bawden *et al.*, 2016) give us the opportunity to critique this method, in particular in terms of the possibility of extending the technique to multiple aspects of speaker identity. Similar experiments on gender adaptation were performed concurrently by van der Wees *et al.* (2016) and Wang *et al.* (2016). Speaker gender is a good starting point for integrating information concerning speaker identity, since it is relatively easy to identify (through knowledge of the person's name if the text is written) and is generally binary. As well as influencing word choice and style, speaker gender is also explicitly visible in languages which display grammatical gender agreement with the speaker. For example, the English sentence "I am happy" would be translated into French "Je suis heureuse" for a female speaker and "Je suis heureux" for a male speaker. When translating from English into French, standard MT systems are limited to selecting the most probable form given the linguistic context, which in most cases is the masculine form "heureux".

We performed a series of simple domain adaptation experiments, consisting of adapting a standard SMT system trained using Moses (Koehn *et al.*, 2007), by modifying both the data used to train the components of the system and the data used for tuning. Our aim was to evaluate the effect on translation performance of training gender-specific language models and gender-specific translation models and of tuning to a gender-specific dataset. The idea of domain adaptation by data partitioning is to annotate the sentences of the data according to the different aspects being studied (the different

*domains* in the broader sense of the term) and to use the data selectively in the training and tuning of the MT system based on these annotations. In our case, the domain was the gender of the speaker, so each dataset was divided into two, corresponding to the sentences spoken by female characters and those spoken by male characters.[2] We used subtitles annotated for speaker gender, taken from the TVD *Big Bang Theory* reproducible corpus (Roy *et al.*, 2014), resulting in approximately 10,000 gender-annotated utterances. The results of the automatic evaluation are summarised in Table 1, using two automatic metrics, BLEU (Papineni *et al.*, 2002) and METEOR (Lavie & Denkowski, 2009).[3] The baseline system is shown in the first row of the table – no specific adaptation is performed and all development data (containing both male and female speakers' utterances) is used to tune the model.

| | | BBT-test$_{male}$ | | BBT-test$_{female}$ | |
|---|---|---|---|---|---|
| Model adaptation | Tuning data | BLEU | METEOR | BLEU | METEOR |
| *(i) Choice of the tuning set* | | | | | |
| ∅ | all | 23.91 | 0.434 | 25.16 | 0.450 |
| ∅ | male | **24.09** | **0.438** | **25.72** | 0.450 |
| ∅ | female | 23.67 | 0.431 | 25.22 | 0.446 |
| *(ii) Addition of a gender-specific language model* | | | | | |
| +LM$_{male}$ | all | **24.17** | **0.436** | 24.80 | 0.447 |
| +LM$_{female}$ | all | 23.35 | 0.430 | 24.13 | 0.443 |
| +LM$_{male}$ | male | 23.92 | 0.435 | 25.39 | 0.448 |
| +LM$_{female}$ | female | 23.97 | 0.444 | **26.25** | **0.459** |
| *(iv) Addition of a gender-specific language model and translation model* | | | | | |
| +LM$_{male}$+TM$_{male}$ | all | 24.06 | 0.434 | 25.36 | 0.449 |
| +LM$_{female}$+TM$_{female}$ | all | 23.60 | 0.431 | 24.55 | 0.444 |
| +LM$_{male}$+TM$_{male}$ | male | **24.18** | **0.436** | **25.69** | **0.451** |
| +LM$_{female}$+TM$_{female}$ | female | 22.64 | 0.422 | 24.91 | 0.441 |

Table 1: Translation performance after adaptation of the phrase-based SMT model to gender. LM$_x$ refers to the addition of a language model trained on data labelled as $x$. Similarly TM$_x$ refers to the addition of a translation model (phase table) trained on data labelled as $x$.

The results show that small gains in translation performance can be obtained from domain adaptation, although these improvements are not significative. Providing a gender-specific tuning set gives gains of +0.1 BLEU and +0.004 METEOR for the male speakers' utterances, and gains of +0.56 BLEU for the female speakers' utterances. Adding a language model that has been trained exclusively on female or male utterances also provides some very slight improvements for both genders, as does the addition of a gender-specific translation model. However a manual evaluation revealed that the slight improvements seen were not due to improvement in grammatical gender agreement at all. Most improvements were thanks to an improved lexical choice when compared to the baseline translation, followed by lexical additions or deletions. It appears that domain adaptation did indeed adapt slightly to the differently gendered datasets, but not necessarily for the reasons we might think; each dataset has its own distinct lexical properties, and it is possible that the improvements were simply due to minor lexical specificities of the two datasets and not necessarily due to a real gender bias. Data sparsity, in particular for female speakers is a problem with this method, which could in part explain why the male-adapted model performs highly on the female-specific test set.

[2]Male utterances were 3.8 times more frequent than female utterances.

[3]BLEU scores are from 0 to 100, with 100 being the highest score, and METEOR are from 0 to 1, 1 being the highest score.

## 2.2 Analysis and perspectives

The main problems found with this method is that data partitioning leads to smaller datasets being used for training and tuning. MT is a domain in which having large amounts of data is a must to produce a robust and high-performing model, so the decision to reduce the size of the data used is not one that should be taken lightly.

It is difficult to imagine this method being extended to integrate other coarse-grained, sentence-level features. In theory, the method is simple, requiring simply annotating data according to the set of features (e.g. sentence type, formality, topic). However, in practice, partitioning of the data depending on the set of class labels, whose number is multiplied each time a new feature is added, would result in smaller and smaller datasets on which to train and tune models, which would inevitably lead to degradation in translation performance. A solution to this problem proposed by Saluja *et al.* (2011), which involves weighting different sentences according to a similarity measure between the test sentence and each training sentence (based on their set of class labels), showed some improvements, but in a limited domain. The use of factored translation models (Koehn & Hoang, 2007) allows more linguistic input to be provided for each of the words, but has had limited impact on translation quality and suffers from computational problems when scaling up to large datasets.

Recently, a new method of performing domain adaptation has been introduced, in the context of neural machine translation (NMT). Sennrich *et al.* (2016) show how it is possible to control for politeness on the sentence-level by introducing an additional feature (*side constraint*) as an arbitrary feature at the beginning of each sentence. The flexibility of this approach appears promising for integrating a range of different sentence-level aspects, including extra-linguistic information related to speaker identity.

# 3 Linguistically motivated cross-lingual pronoun prediction

Other forms of context necessarily require a more sophisticated representation of the linguistic properties of discourse. One of the most studied contextual aspects of discourse is coreference and the translation of pronouns, particularly between languages that do not have the same system of grammatical gender (Le Nagard & Koehn, 2010; Guillou, 2016). Take for example the language pair English-French. French common nouns and pronouns are marked for grammatical gender, whereas the English pronouns *it* and *they* are not. In order to correct translate the pronoun *it* in the English sentence "it was blue", it is necessary to know the grammatical gender of the French coreferential antecedent of the pronoun. For example, if *it* refers to a box (*une boîte*), the correct translation is the feminine pronoun *elle*, whereas if *it* refers to a toe (*un orteil*), the correction translation is the masculine pronoun *il*. There is also a third option (which can be used in certain contexts) to translate *it* using the gender-neutral demonstrative *ce*.

The problem of ensuring accurate pronominal translation has received much interest in the discourse in MT community and a shared task has been organised for the past few years, dedicated to providing solutions to the problem (e.g. Guillou *et al.*, 2016). The task is a cross-lingual pronoun prediction task, for which participating classification systems aim to correctly predict target pronoun forms in the target sentence, based on contextual information in the source and target sentences. We focus on one of the language directions proposed by the 2016 task: English into French, for which the aim was to predict the correct French translation of the English subject pronouns *it* and *they*, from the set of possible classes *il*, *ils*, *elle*, *elles*, *on*, *ce*, *cela* and *OTHER*. An example of the data provided for the

task is shown in Figure 3. Target sentences were provided as lemma-tag pairs, rather than as surface forms, to avoid systems relying solely on the grammatical gender of the immediately surrounded words to determine gender, which is an unrealistic scenario in a real translation setting.[4]

| Pronoun class | Lemma+tag | English source sentence | French target sentence | Word alignment |
|---|---|---|---|---|
| ils | il\|PRON | They 're extremely costly . | REPLACE_0 être\|VER très\|ADV coûteux\|ADJ .\|. | 0-0 1-1 2-2 3-3 4-4 |

Figure 1: An example of pronoun prediction task data at WMT 2016. The first column represents the class label to be predicted and the second column the lemma and tag of the pronoun. The position of the pronoun to be predicted is marked by the placeholder REPLACE_0 in this example.

A variety of different strategies were used by task participants, who relied on different amounts of contextual, linguistic information. The winning system (Luotolahti *et al.*, 2016) was a stacked recurrent neural network (RNN) system, which did not rely on any other contextual information than the information provided (as shown in Figure 3) and did not look beyond sentence boundaries. A second neural network approach also scored highly in the task: Dabre *et al.* (2016) used a simple RNN architecture with an attention mechanism and trained only using IWSLT data (whereas the winning system was trained on all data provided). Many of the other participants, including the current authors, chose to concentrate on the integration of more linguistic information into simpler (mainly linear) classification systems. Many participants used part-of-speech (PoS)-tagging, parsing and coreference chains. Stymne's (2016) system, a linear classifier using multiple linguistic annotations, was ranked second out of nine systems, showing that simple linear classifiers can perform just as highly as more sophisticated architectures such as neural network systems if good quality linguistic information is used.

## 3.1 Experiments and results

Our system (Bawden, 2016), followed this second strategy of including as much relevant linguistic knowledge into the system as possible. One of the main aims of our submission was to evaluate the capacity of linguistic tools and resources to provide accurate linguistic knowledge that, if perfect, should be sufficient to predict the correct pronoun. We used a variety of linguistic annotations and heuristics to provide features to a random forest classifier, implemented in Scikit-learn (Pedregosa *et al.*, 2011), relying heavily on linguistic tools and external sources. As source-side information, we used automatic PoS-tagging, dependency parsing and coreference resolution annotations, provided by the Stanford toolkit (Manning *et al.*, 2014). On the target-side, we used dependency parsing using the Mate Parser (Bohnet & Nivre, 2012) and a parse model trained on lemmas and morphological information found in the morphological and syntactic lexicon, the Le*fff* (Sagot, 2010). We also integrated the prediction provided by the language model baseline (Tiedemann, 2016) and identified local windows of syntactic and morphological patterns that were particularly linked to certain classes.[5]

We chose to explicitly perform coreference resolution, using the Stanford toolkit. Since the tool is not available for French, we performed coreference resolution on the English source sentence and

[4]In the 2015 version of the task in which tokens were used on the target side, not a single system beat the baseline system, which was a simple $n$-gram language-model (Hardmeier *et al.*, 2015). The use of just lemmas in the 2016 task encouraged participants to better model the context and use richer information than just surface forms.

[5]These final features were particularly useful for predicting impersonal *il* pronouns, as well as anomalies in the training data, where the English pronoun was not translated by a French pronoun, and the gold class was therefore the class OTHER.

| | Classified as | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ce | elle | elles | il | ils | cela | on | other | SUM | P (%) | R (%) | F (%) |
| ce | **54** | 1 | 0 | 11 | 0 | 0 | 0 | 2 | 68 | 91.53 | 79.41 | 85.04 |
| elle | 0 | **13** | 1 | 6 | 0 | 2 | 0 | 1 | 23 | 41.94 | 56.52 | 48.15 |
| elles | 1 | 2 | **3** | 1 | 13 | 1 | 0 | 4 | 25 | 23.08 | 12.00 | 15.79 |
| il | 2 | 7 | 0 | **44** | 1 | 2 | 1 | 4 | 61 | 61.97 | 72.13 | 66.67 |
| ils | 0 | 1 | 9 | 0 | **56** | 0 | 0 | 5 | 71 | 75.68 | 78.87 | 77.24 |
| cela | 0 | 5 | 0 | 7 | 0 | **13** | 1 | 5 | 31 | 72.22 | 41.94 | 53.06 |
| on | 0 | 0 | 0 | 0 | 2 | 0 | **5** | 2 | 9 | 55.56 | 55.56 | 55.56 |
| OTHER | 2 | 2 | 0 | 2 | 2 | 0 | 2 | **75** | 85 | 76.53 | 88.24 | 81.97 |
| SUM | 59 | 31 | 13 | 71 | 74 | 18 | 9 | 98 | | | | |
| Micro-averaged | | | | | | | | | | 70.51 | 70.51 | 70.51 |
| Macro-averaged | | | | | | | | | | 62.31 | **60.58** | 60.43 |

Table 2: A decomposition of results on the test set for our submission to the WMT 2016 cross-lingual pronoun prediction task. The results are slightly higher than those reported in the official scoreboard due to the resolution of a minor bug found after the submission deadline.

automatically transferred the information through the alignments to the French target sentence. As shown in Figure 2, (i) we used the automatic alignments to find the pronoun to which the placeholder is aligned, (ii) we then use the English coreference chains to identify the antecedent of the English pronoun, and (iii) we again use the automatic alignment to identify the French antecedent of the pronoun. The gender of the pronoun is given by the French antecedent (found in the lexicon) and number is provided by the number of the English aligned pronoun. Both of these values were used alongside the other features in our classifier.

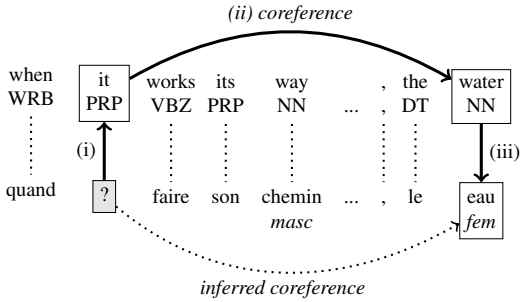



Figure 2: Use of coreference chains to determine gender and number of anaphoric (or, as shown here, cataphoric) pronouns.

The breakdown of the results of our system is shown in Table 2. The official metric for the task was macro-averaged recall, which means that more importance is given to rarer classes, such as *elle*, *elles* and *on*. The system scores ranged from 36.35 to 65.70, with the baseline system at 50.85. The overall score of our system 59.32 put us at sixth position out of nine systems (excluding the baseline system). Our system scored +8.47 points higher than the baseline, showing that exploiting contextual linguistic information is useful for the task.

### 3.2 Analysis and perspectives

As with all systems submitted to the task, the most difficult classes to predict were the rarest classes: *elle*, *elles* and *on*. In terms of pronouns determined by coreference, where the neural network systems were limited by sentence boundaries, the linguistically-motivated systems were limited by the performance of the tools providing the linguistic annotation. Our analysis of the performance of the Stanford Coreference tool showed that a correct antecedent was provided in only 52.5% of cases and 32% of pronouns were linked only to other pronouns rather than to full noun phrases, even when searching for the noun phrase by transitivity. The tool also fails to identify impersonal uses of pronouns (supplying coreference chains for 18 impersonal pronouns out of 25). Whilst resources and linguistics tools can provide useful information for integration such contextual information, it is important not underestimate their lack of robustness in certain contexts. Neural classification methods, with the use of richer representations of both source and target sentences through embeddings, could provide a good setting for integrating this contextual information more effectively, without having to rely as much on the accuracy of the annotation tools.

## 4 Stylistic choices: the case of tag questions

The final focus of this paper is on an even more complicated notion of context, which is more difficult to characterise than the two previous types of contexts discussed in the two previous sections. Unlike speaker identity or grammatical gender in coreference chains, which are deterministic and easily discernible, we focus here on one particular aspect of stylistic choice, the English tag question. Common in spoken English (particularly British English), the tag question, which is best known in its canonical form, formed of an auxiliary verb and a pronoun of the form *isn't it?*, *can we?* or *would you?*, etc., is a much studied phenomenon in the field of linguistics (McGregor, 1995; Kimps, 2007). According to the Longman Grammar (Biber *et al.*, 1999), approximately 20% of the questions found in the conversational part of the corpus they analysed were tag questions.

Question tags are peripheral interrogative elements, typically appended to a declarative sentence, with the effect of modifying the sentence, to express a range of different attitudes, including doubt, surprise, contempt, etc. or simply to facilitate the flow of conversation. There are two main types of tag question, grammatical tag questions, also known as canonical tag questions, which are grammatically bound to the main anchor clause, for example:

(1) *You do believe in happy endings*, **don't you?**

(2) *He can't do that*, **can he?**

However, there also exist lexical (or invariant) tag questions, which are not grammatical bound in the same way, and are in fact the most frequent form of tag question, also commonly appearing in languages other than English. For example:

(7) *He's a proper bad man*, **innit?**

(8) *There's got to be a cure*, **right?**

The function of tag questions is complex and multi-facetted. They have often been referred to as dialogue facilitators, as ways of "keeping the conversation going" and "inviting listeners to

communicate" (Soars & Soars, 2000). However they have also been analysed as being modalisers to the main proposition of the utterance, modifying it to express the speaker's attitude to the proposition (McGregor, 1995). This bias based on the speaker's belief in the truth of the initial proposition is also linked to the understanding and common knowledge between speaker and interlocutor. The choice of question tag can have an important impact on the meaning of an utterance, in particular concerning speaker certainty, politeness and conduciveness. For example, a "they're a bit strange, aren't they?" suggests that the speaker thinks them to be strange, but demands confirmation from the listener in an inviting way, whereas "they're a bit strange, are they?" suggests that the listener has stated they are a bit strange and the speaker is questioning this, potentially in a slightly aggressive manner.

Tag questions are notably difficult to translate automatically because there are no direct equivalents to the English system of tag questions in other languages, with respect to canonical (grammatical) tag questions of the form *isn't it?*, *won't you?*.[6] The form of tag questions may be considered relatively easy to predict, once an utterance is known to be a tag question, because the grammatical structure of a tag question in most cases echoes the auxiliary verb of the main clause. However, there is also the option to use a lexical tag question (e.g. *right?*, *see?*), which complicates this decision. The most difficult decision is to know when to produce a tag question when translating from a language that uses different strategies for the same communicative purpose. SMT systems, which only use a very notion of local, immediate context, often struggle to provide a grammatical and coherent tag question. NMT systems theoretically far better, but the task of generating a tag question where one does not appear in the source sentence is still a considerable challenge.

## 4.1 Predicting the use of English tag questions: presence and form

To our knowledge, ours is the first work on the MT of tag questions in English, despite the vast linguistic literature on the topic. For the language direction French to English, we aim to improve the MT output of sentences whose English reference translations correspond to English tag questions, without degrading translation performance overall. We formalise the problem as a post-edition, classification problem on top of a strong SMT baseline. Given a high-performing phrase-based system (Koehn *et al.*, 2007), trained on parliamentary data, conference talks and subtitles, we modify the sentences classified as containing a tag question by appending the tag question form predicted by the classifier.

**Tag question identification** We first automatically annotate a subset of the OpenSubtitle parallel corpus (Lison & Tiedemann, 2016) for the presence or absence of tag questions on the English side of the corpus and for the form of the question tag if the subtitle is indeed a tag question. We use a sequence of heuristic, lexical-based rules based on our knowledge of English syntax and frequent patterns. A manual evaluation on a 500-subtitle subset of our annotations shows that precision and recall are approximately 98% and 100% respectively for the identification of sentence-final grammatical tag questions whose anchor clause is in the same subtitle. We divide the corpus consecutively into four datasets (train, dev1, dev2 and test), whose sizes are shown in Table 3. The number of different tag questions are also provided (corresponding to approximately 1% of sentences in each dataset). The distribution between the different question tag forms is very unequal (with the form "right" making up 20% of all tag question forms. This makes both learning and evaluation difficult.

[6] Grammatical tag questions do appear in some other languages, but this is rare (Axelsson, 2011).

| | #sents | #TQs all | grammatical | lexical |
|---|---|---|---|---|
| train | 18.5M | 162,124 (0.9%) | 71,889 (0.4%) | 90,235 (0.5%) |
| dev1 | 5M | 49,908 (1%) | 15,070 (0.3%) | 34,838 (0.7%) |
| dev2 | 5M | 48,825 (1%) | 13,931 (0.3%) | 34,894 (0.7%) |
| test | 5M | 48,676 (1%) | 13,212 (0.3%) | 35,646 (0.7%) |

Table 3: Distribution of English TQs in the four datasets.

**Experimental setup** We perform classification in two separate steps, first by predicting whether or not a tag question should be used in the target translation and secondly by predicting the true form of the question tag (e.g. *isn't it*, *don't you think*) for those sentences predicted to be tag questions by the first classifier. Both classifiers are linear classifiers trained using Vowpal Wabbit (Langford *et al.*, 2009).

The first classifier, whose task is to predict the presence or absence of a tag question, is trained on features extracted only from source data from the train set. The second classifier uses information from the current and following source sentences, the current and following target sentence and the score from the first classifier. Only those sentences to which the first classifier assigns a score above a certain threshold are sent to the second classifier, all other sentences being considered as non-tag questions. The value of this threshold at training time (which sentences are selected for inclusion into the second classifier's training set) and at testing time (which sentences will be eligible for being assigned a tag question by the second classifier) constitute the two hyperparameters of our system. The first classifier is trained on the training set, and the second classifier is trained on the subpart of the dev1 set selected by the first hyperparameter, both hyperparameters being optimised on the dev2 set. Features include significant bag of word features (unigram, bigram and trigram) filtered using a $G^2$ test, the $n$-grams of the sentence, the presence of a lexical tag in French and the presence of an affirmative or negative reply in the following subtitle.

| | Step 1: TQ identification | Step 2: TQ labelling precision (%) | | | BLEU scores | |
|---|---|---|---|---|---|---|
| | F-score | ALL | TQ | Non-TQ | TQ | ALL |
| Baseline | 48.3 | 99.02 | 29.56 | 99.7 | 31.26 | 34.11 |
| Our system | 62.6 | 99.15 | 20.94 | 99.9 | 32.08 | 34.12 |
| Topline | - | 100 | 100 | 100 | 46.68 | 34.33 |

Table 4: Post-edition results on the test set (f-score, precision and BLEU). Scores marked "TQ" apply only to sentences which contain a tag question in the reference translation. Scores marked "ALL" apply to all sentences. The baseline score is the machine translated output and the topline is the machine translated output, post-edited with the gold question tags.

**Classification Results** Results are shown in Table 4 for both classification steps. For the second step, we provide a total classification precision on all 214 labels (including the label *none* when no tag question is present).[7] However we also provide separate labelling precision for gold tag questions

[7]The second classifier is trained on the predicted outputs of the first classifier and so it is necessary to include the "none" label in the second step to allow the system to go back on the decision of the first classifier.

and for gold non-tag questions to give more insight into the performance of the classifier.[8] Finally, translation performance is evaluated using the automatic evaluation metric BLEU (Papineni *et al.*, 2002). In these preliminary experiments, we see that the overall labelling precision increases with respect to the baseline, but this is chiefly due to an increased precision in predicting when a subtitle is not a tag question. Our system predicts far more *none* labels than the baseline system, resulting in a lower precision for gold TQs. First observations show that our system has higher precision on those subtitles predicted as tag questions than the baseline, but this is at the expense of a lower recall (as reflected in the precision on gold TQs). We see slight increases in BLEU score for the entire corpus and on gold TQs. The lack of correlation between TQ precision and TQ BLEU should be investigated, and is possibly due to the fact that wrong question tag labels are penalised more than an absent tag label. Further experiments and a detailed manual evaluation would have to confirm this. The main problem with evaluating such a subjective, stylistic aspect is that the automatic evaluation metrics used here do not take into account the fact that several correct tag questions could be possible for a given sentence.

### 4.2 Analysis and perspectives

In future work, we aim to build on these preliminary experiments to improve the classification process and consequently the translation of tag questions. Possible perspectives include combining statistical classification for the identification step with robust rules for the prediction of the grammatical tag forms, which are most often very predictable given a main sentence.

Although tag question prediction may appear a very minor aspect of translation, appearing in only 1% of sentences, correctly predicting the tag question form could greatly improve the fluidity, coherence and naturalness of translated dialogues. In this experiment, we identified tag questions in the English subtitles using carefully constructed rules, but a future step could be to automatically identify such phenomena, which are difficult to translate, using the knowledge gained through a linguistic analysis of our corpus and an in-depth comparison of reference and baseline translations.

## 5 Conclusion

Context, in its many different forms, is crucial when translating speech-like texts. We have shown, through three different experiments on three aspects of contextual information, how integrating context into the translation process can help translation performance, whether during decoding as with our first experiment on domain adaptation or as a post-edition step as with the last two experiments. However, these methods remain limited. Domain adaptation by data partitioning restricts the amount of data used, pronoun predicting using external resources is, at present, limited by the lacunas in automatic coreference resolution, and we are far from being able to model the many different aspects of style, which, like tag questions, can greatly impact translation quality. The arrival of neural learning methods, which provide a promising architecture for taking into account richer linguistic information (Sennrich & Haddow, 2016), pave the way for integrating rich, contextual information into MT systems, and importantly, could provide a unified approach for the integration of the different types of context.

[8] TQ precision is the percentage of reference tag questions that are correctly labelled with the exact form of the tag question. Non-TQ precision is the percentage of reference non-tag questions that are correctly labelled.

# References


AXELROD A., HE X. & GAO J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, p. 355–362, Edinburgh, UK.

AXELSSON K. (2011). A cross-linguistic study of grammatically-dependent question tags. *Studies in Language*, **35**(4), 793–851.

BAWDEN R. (2016). Cross-lingual Pronoun Prediction with Linguistically Informed Features. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 564–570, Berlin, Germany.

BAWDEN R., WISNIEWSKI G. & MAYNARD H. (2016). Investigating gender adaptation for speech translation. In *Proc. of the 23rd Conférence sur le Traitement Automatique des Langues Naturelles*, TALN'16, p. 490–497, Paris, France.

BIBER D., JOHNANSSON S., LEECH G. & QUIRK R. (1999). *Longman Grammar of Spoken and Written English*. Longman, London.

BOHNET B. & NIVRE J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, p. 1455–1465, Jeju Island, Korea.

COATES J. (1986). *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. London: Longman.

DABRE R., PUZIKOV Y., CROMIERES F. & KUROHASHI S. (2016). The Kyoto University Cross-Lingual Pronoun Translation System. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 571–575, Berlin, Germany.

FINCH A., SUMITA E. & NAKAMURA S. (2009). Class-Dependent Modeling for Dialog Translation. *IEICE Transactions on Information and Systems*, **92**(12), 2469–2477.

FOSTER G. & KUHN R. (2007). Mixture-Model Adaptation for SMT. In *Proc. of the 2nd Workshop on Statistical Machine Translation*, WMT'07, p. 128–135, Prague, Czech Republic.

GUILLOU L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, School of Informatics. University of Edinburgh.

GUILLOU L., HARDMEIER C., NAKOV P., STYMNE S., TIEDEMANN J., VERSLEY Y., CETTOLO M., WEBBER B. & POPESCU-BELIS A. (2016). Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 525–542, Berlin, Germany.

HARDMEIER C. (2012). *Discourse in statistical machine translation. a survey and a case study*. PhD thesis, Uppsala University, Uppsala, Sweden.

HARDMEIER C., NAKOV P., STYMNE S., TIEDEMANN J., VERSLEY Y. & CETTOLO M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proc. of the 2nd Workshop on Discourse in Machine Translation*, DiscoMT'15, p. 1–16, Lisbon, Portugal.

HOVY D. (2015). Demographic Factors Improve Classification Performance. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP'15, p. 752–762, Beijing, China.

KIMPS D. (2007). Declarative constant polarity tag questions: A data-driven analysis of their form, meaning and attitudinal uses. *Journal of Pragmatics*, p. 270–291.

KOEHN P. & HOANG H. (2007). Factored translation models. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 868–876, Prague, Czech Republic.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL'07, p. 177–180, Prague, Czech Republic.

KUHN R., NGUYEN P., JUNQUA J.-C., GOLDWASSER L., NIEDZIELSKI N., FINCKE S. & CONTOLINI M. (1998). Eigenvoices for speaker adaptation. In *Proc. of the 5th International Conference on Spoken Language Processing*, IWSLT'98, p. 1771–1774, Sydney, Australia.

LAKOFF R. (1975). *Language and woman's place: text and commentaries*. New York: Oxford University Press.

LANGFORD J., LI L. & ZHANG T. (2009). Sparse Online Learning via Truncated Gradient. *The Journal of Machine Learning Research*, p. 777–801.

LAVIE A. & DENKOWSKI M. J. (2009). The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, **23**(2-3), 105–115.

LE NAGARD R. & KOEHN P. (2010). Aiding pronoun translation with co-reference resolution. In *Proc. of the 5th Workshop on Statistical Machine Translation*, WMT'10, p. 252–261, Uppsala, Sweden.

LISON P. & TIEDEMANN J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proc. of the 10th Language Resources and Evaluation Conference*, LREC'16, p. 923–929, Portorož, Slovenia.

LUOTOLAHTI J., KANERVA J. & GINTER F. (2016). Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 596–601, Berlin, Germany.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL'14, p. 55–60, Baltimore, USA.

MCGREGOR W. (1995). The English 'tag question': A new analysis, is(n't) it? *On Subject and theme: A discourse functional perspective*, **118**, 91–121.

MIRKIN S., NOWSON S., BRUN C. & PEREZ J. (2015). Motivating Personality-aware Machine Translation. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, p. 1102–1108, Beijing, China.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, p. 311–318, Philadelphia, USA.

PECINA P., TORAL A. & VAN GENABITH J. (2012). Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. In *Proc. of the 24th International Conference on Computational Linguistics*, COLING'12, p. 2209–2224, Mumbai, India.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

ROY A., GUINAUDEAU C., BREDIN H. & BARRAS C. (2014). TVD: A Reproducible and Multiply Aligned TV Series Dataset. In *Proc. of the 9th Language Resources and Evaluation Conference*, LREC'14, p. 418–425, Reykjavik, Iceland.

SAGOT B. (2010). The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proc. of the 7th International Conference on Language Resources and Evaluation*, LREC'10, p. 2744–2751, Valletta, Malta.

SALUJA A., LANE I. & ZHANG Y. (2011). Context-aware Language Modeling for Conversational Speech Translation. In *Proc. of the 13th Machine Translation Summit*, p. 97–104, Xiamen, China.

SENNRICH R. & HADDOW B. (2016). Linguistic Input Features Improve Neural Machine Translation. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 83–91, Berlin, Germany.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'16, p. 35–40, California, USA.

SOARS J. & SOARS L. (2000). *New headway English course. Pre-intermediate student's book*. Oxford: Oxford University Press, 2nd edition.

STYMNE S. (2016). Feature exploration for cross-lingual pronoun prediction. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 609–615, Berlin, Germany.

TIEDEMANN J. (2016). A Linear Baseline Classifier for Cross-Lingual Pronoun Prediction. In *Proc. of the 1st Conference on Machine Translation*, p. 616–619, Berlin, Germany.

VAN DER WEES M., BISAZZA A. & MONZ C. (2016). Measuring the Effect of Conversational Aspects on Machine Translation Quality. In *Proc. of the 26th International Conference on Computational Linguistics*, COLING'16, p. 2571–2581, Osaka, Japan.

VOLKOVA S., WILSON T. & YAROWSKY D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, p. 752–762, Beijing, China.

WANG L., ZHANG X., TUY Z., WAY A. & LIU Q. (2016). Automatic construction of discourse corpora for dialogue translation. In *Proc. of the 10th International Conference on Language Resources and Evaluation*, LREC'16, p. 2748–2754, Portorož, Slovenia.

# Construction de lexiques pour l'extraction des mentions de maladies dans les forums de santé

Elise Bigeard[1, 2]
(1) laboratoire STL, Lille 3
(2) Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France
elise.bigeard@u-bordeaux.fr

RÉSUMÉ

Les forums de discussion et les réseaux sociaux sont des sources potentielles de différents types d'information, qui ne sont en général pas accessibles par ailleurs. Par exemple, dans les forums de santé, il est possible de trouver les informations sur les habitudes et le mode de vie des personnes. Ces informations sont rarement partagées avec les médecins. Il est donc possible de se fonder sur ces informations pour évaluer les pratiques réelles des patients. Il s'agit cependant d'une source d'information difficile à traiter, essentiellement à cause des spécificités linguistiques qu'elle présente. Si une première étape pour l'exploration des forums consiste à indexer les termes médicaux présents dans les messages avec des concepts issus de terminologies médicales, cela s'avère extrêmement compliqué car les formulations des patients sont très différentes des terminologies officielles. Nous proposons une méthode permettant de créer et enrichir des lexiques de termes et expressions désignant une maladie ou un trouble, avec un intérêt particulier pour les troubles de l'humeur. Nous utilisons des ressources existantes ainsi que des méthodes non supervisées. Les ressources construites dans le cadre du travail nous permettent d'améliorer la détection de messages pertinents.

ABSTRACT

**Building lexica for extraction of mentions of diseases from healthcare fora**

Discussion fora and social networks potentially provide various types of information, which usually are not available in other sources. For instance, it is possible to find in discussion fora information on habits and lifestyle of persons, while this information is seldom shared with medical doctors. Hence, it is possible to rely on this information for evaluating the real habits of patients. Yet, this information source is difficult to process, mainly because it offers some linguistic specificities. If the first step for exploring discussion fora consists of indexing the medical terms occurring in the messages with concepts from medical terminologies, this task is very difficult because patient utterances are different from official terminologies. We propose a method that permits to create and enrich lexica with terms and expressions meaning disorders or abnormalities, with a special interest to mood disorders. We exploit existing resources and non-supervised methods. Resources built during the work allow to improve the detection of relevant messages.

MOTS-CLÉS : Traitement automatique du langage, forums de discussion, extraction, terminologies, langage patient, maladies.

KEYWORDS: Natural Language Processing, Discussion Forums, extraction, patient language, disorders.

# 1 Introduction

Nous nous intéressons à l'usage réel qui est fait des médicaments par les patients et recherchons des méthodes pour détecter des mésusages de médicaments fréquents. Il est rare qu'un patient en parle au médecin qui lui a fait la prescription, soit parce qu'il n'est pas conscient d'utiliser le médicament dans des conditions non prévues (présence d'une contre-indication, d'une intéraction avec un autre produit, une erreur de dose ou de voie d'administration etc.), soit parce que ce mésusage est volontaire pour obtenir un effet particulier. Pour en savoir plus, il est nécessaire de se tourner vers d'autres sources d'information, telles que les forums de discussion. En effet, dans l'anonymat et sans avoir besoin de fournir d'effort particulier, les patients discutent volontiers de leurs maladies et de leurs pratiques sur les forums, nous ouvrant ainsi une fenêtre sur leur vie et leurs actions (Gauducheau, 2008). Une première étape pour aller vers l'exploration de forums consiste typiquement à indexer les termes médicaux correspondant aux maladies présentes dans les messages avec des termes de terminologies médicales de référence. La difficulté principale est que ces termes ont des formulations très différentes et diverses dans le langage patient.

Par exemple, pour désigner les troubles de l'humeur, des termes standards tels que *dépression* ou *déprimé* côtoient des expressions voisines comme dans cette phrase : *Je ne supporte rien je suis à fleur de peau c'est horrible, je suis hyper nerveuse et obligée de compléter avec une benzo pour me calmer tellement je suis dans un état de nerfs prononcé.* Nous pouvons ainsi observer plusieurs expressions de patients, qui ne peuvent pas être trouvées dans les terminologies ou ontologies.

D'autres obstacles, tels les fautes d'orthographe ou les abréviations non-standard, sont spécifiques au médium des réseaux sociaux, et même parfois à un réseau social particulier.

Pour résoudre ces difficultés, nous proposons une méthode permettant de créer et enrichir des lexiques avec des termes et expressions désignant une maladie. Nous portons un intérêt particulier aux troubles de l'humeur.

La constitution de lexiques est une tâche du TAL explorée depuis de nombreuses années. Nous nous intéressons ici aux lexiques qui associent une entrée à une liste de mots sémantiquement proches. La proximité sémantique peut référer à des mesures variées, dont notamment les relations synonymiques et paradigmatiques. (Budanitsky & Hirst, 2006) (Adam *et al.*, 2013).

La plupart des ressources sémantiques existantes se limitent aux relations synonymiques. Elles ne sont pas forcément accompagnées d'une mesure de proximité ou de confiance, rendant difficile la distinction entre synonymes, quasi-synonymes ou relations d'un autre type. Une autre limitation courante des ressources externes concerne le registre. Le vocabulaire utilisé par les usagers sur les forums de discussion ne coïncide pas avec le vocabulaire trouvé dans les terminologies de spécialité, plutôt destinées aux professionnels d'un secteur.

Pour le domaine médical, sur l'exemple de l'initiative Consumer Health Vocabularies (Zeng & Tse, 2006) en anglais, des initiatives existent pour créer des vocabulaires patient en français : sur le cancer du sein (Eholié *et al.*, 2016; Messai *et al.*, 2006), ou pour le domaine médical en général (Grabar & Hamon, 2014). Mais ces travaux ne couvrent que peu de termes et sont rarement adaptés au média particulier des forums de discussion.

Le crowdsourcing peut être une autre approche, mais est limité par la motivation et l'expertise des participants (Khare *et al.*, 2016).

Outre des ressources externes, il est possible de se fonder sur des méthodes distributionnelles pour

faire émerger du corpus les relations sémantiques qui nous intéressent. Ces méthodes utilisent les relations paradigmatiques comme dimension de la proximité sémantique entre des termes. Ce sont des méthodes non supervisées, capables de tirer parti de gros corpus et de créer des ressources fournies. De nombreux travaux récents s'intéressent à ces méthodes pour traiter des données médicales (Choy *et al.*, 2016) (Liu *et al.*, 2016).

Nous proposons d'explorer les ressources existantes, créées par spécialistes ou le crowdsourcing, et de construire de nouvelles ressources issues de nos corpus en exploitant des méthodes non supervisées. Ces ressources vont nous permettre de détecter les messages relatifs à un trouble donné.

# 2 Ressources

## 2.1 Forums de discussion

Notre corpus provient du forum de santé *www.doctissimo.fr* et de sa section dédiée aux anti-dépresseurs [1]. Les utilisateurs y parlent de leurs troubles de l'humeur, mais pas nécessairement d'anti-dépresseurs ou de dépression. Nous avons collecté les messages postés de 2010 à 2015. Dans chaque sujet, nous avons conservé uniquement le premier message car nous avons estimé que des informations à propos des actions d'un utilisateur sont plus susceptibles d'y apparaître, tandis que les réponses contiennent le plus souvent des opinions et des réactions à ce premier message. Notre corpus contient 6 185 posts composés de 828 250 occurrences de mots. Chaque message contient généralement quelques phrases. Ce corpus est utilisé pour détecter des informations sur l'usage des médicaments, et pour analyser le comportement des patients.

## 2.2 Ensemble de référence avec les noms de maladies

Un ensemble de 23 noms de maladies est extrait des Résumés de Caractéristiques du Produit (RCP) publiés par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM). Il s'agit d'une ressource indépendante de notre travail. Pour permettre sa commercialisation, chaque médicament doit recevoir une autorisation de mise sur le marché (AMM) dont le RCP est une annexe : elle constitue par la suite un document de référence qui contient toutes les informations plus particulièrement destinées aux professionnels de santé. Les termes de notre ensemble sont liés aux maladies pour lesquelles sont prescrits des antidépresseurs, anxyolitiques et autres médicaments utilisés pour traiter les troubles de l'humeur. Cette liste contient des termes tels que *dépression*, *anxiété*, *nerveux*, *phobie*, *panique* ou *angoisse*. Chaque terme est associé à un code CIM10 (OMS, 1995), comme par exemple *anxiété / F41.9* ou *dépression / F32*. Le choix de la CIM-10 plutôt qu'une autre classification a été fait par des experts. Notre objectif est de trouver les messages dans lesquels ces termes, ou leurs équivalents sémantiques, sont présents. Dans la suite de ce papier, ces 23 termes seront appelés *seeds* lorsqu'ils sont utilisés comme source pour créer ou ajuster des ressources sémantiques.

Ces 23 seeds sont associées à deux jeux de codes CIM10. Le premier, appelé *seeds1*, correspond aux codes tels qu'extraits automatiquement des RCP. Dans ce cas, chaque code correspond à un diagnostic distinct. Cependant, les patients ne font pas forcément la distinction entre certains diagnostics

1. *http ://forum.doctissimo.fr/medicaments/antidepresseurs-anxiolytiques/liste sujet-1.htm*

proches et ne marquent pas cette distinction dans leur façon de parler des maladies. En effet, si nous recherchons les 30 termes les plus proches de chaque seed en exploitant les sorties de l'algorithme Word2Vec (Mikolov *et al.*, 2013a,b) (voir section 3.2 pour sa description) nous pouvons observer que les seeds comme *phobie* et *agoraphobie* ont 38 % de mots en commun, *angoisse* et *panique* 35 % de mots en commun, *stress* et *angoisse* ainsi que *phobie* et *panique* partagent 26 % de leurs mots. Il s'agit donc de termes qui apportent potentiellement des ambiguïtés lors de leur utilisation par les patients.

Nous avons donc créé un second jeu de codes, plus généraux et appelés *seeds2*, où les maladies le plus souvent associées et confondues par les patients figurent sous un même code. Ainsi, *agoraphobie/F40.0* est rapproché du terme moins spécifique *phobie/F40*. *angoisse/F41.0*, *anxiété/F41.9* et *anxiété généralisée/F41.1* sont également convertis vers un code plus général *autres troubles de l'anxiété/F41*. Ce regroupement est effectué à la main par un expert et se fonde d'une part sur les recoupements détectés par Word2Vec et notre connaissance du domaine et du corpus, afin de déterminer quelles seeds devraient être fusionnées.

## 2.3 Messages annotés manuellement

Pour constituer le corpus d'évaluation, 100 messages ont été sélectionnés aléatoirement. Ont été exclus certains messages très longs et peu cohérents car difficiles à annoter et peu représentatifs. Ces 100 messages ont été annotés manuellement par un expert, dont la tâche était d'y associer une expression d'un trouble de l'humeur en relation avec la liste de seeds. Ces annotations ont ensuite été converties en codes *seeds1* et *seeds2*. L'annotation a été effectuée au niveau des phrases. Ces 100 messages constituent les données de référence par rapport auxquelles nous évaluons les résultats produits dans ce travail. L'annotation s'est révélée difficile pour certains codes seeds1 proches. C'est pourquoi seulement 100 posts ont été annotés. Pour des travaux futurs, nous nous baserons uniquement sur seeds2 pour faciliter l'annotation et produire un corpus annoté plus fourni.// Ci-dessous des exemples de phrases et de leurs annotations.

| | *annotation* | phrase |
|---|---|---|
| 1. | *angoisse, dépression* | Bonjour, J'ai besoin de conseils car je ne suis pas bien en ce moment : angoissée en permanence et idées noires très présentes. |
| 2. | *dépression* | Salut Je repose la question mais pour un nouvelle ad que je prends. |
| 3. | *dépression* | J'ai vraiment du mal à me tirer du lit le matin, même après trois semaines. |
| 4. | *phobie* | ça part ces obsession de peur d'avoir peur ? |
| 5. | *angoisse* | pcq mon medecin a dit que mon mal de centre c de l'angoisse qu'en penser vous ? ? |
| 6. | *boulimie* | Bonjour à tous, Je viens de suivre un traitement au Séroplex pendant un an, qui m'a été moralement bénéfique certes, mais dont je paie le prix fort : Plus de 15 kilos en plus, sachant que je faisais 65 kilos il y'a un peu moins d'un an ! ! ! |

Le but de ces annotations n'est pas de poser un diagnostic, mais d'annoter les termes faisant référence à un trouble. Ainsi, dans la dernière phrase présentée ci-dessus, la mention de prise de poids a été annotée en *boulimie*, bien que le reste de la phrase ne semble pas pointer vers un diagnostic de boulimie. Dans les messages non standards, il est rare que le nom de la maladie soit mal orthographié (phrase 5), en revanche les abréviations y sont fréquentes (phrase 2). Parfois, un terme n'est pas assez

précis pour être associé à un trouble donné. Dans ce cas, un choix arbitraire mais consistant a été fait par l'annotateur. C'est par exemple le cas d'idées noires qui est toujours annoté comme dépression. La phrase 3 donne un exemple de phrase où il est difficile d'isoler un terme univoquement associé à une maladie.

# 3 Méthodes

## 3.1 Prétraitement du corpus

Les messages sont segmentés en phrases et mots, étiquetés morpho-syntaxiquement et lemmatisés par Treetagger (Schmid, 1994). Les chiffres sont remplacés par une marque de substitution. Les diacritiques et la casse sont neutralisés pour diminuer la variation orthographique. Aucune correction orthographique n'est effectuée. Les mots vides peuvent avoir une importance pour les méthodes exploitées et n'ont donc pas été supprimés.

## 3.2 Construction des ressources sémantiques

La construction des ressources sémantiques est l'étape principale de notre méthode. La tâche consiste à créer des lexiques constitués d'expressions proches des troubles étudiés. Pour réaliser cette tâche, nous explorons deux approches : (1) utilisation de ressources existantes (*Lexique.org, Wikidata* et *JeuxdeMots*) desquelles les expressions pertinentes pour notre travail seront extraites ; (2) exploitation d'approches basées sur le corpus, en l'occurrence de méthodes distributionnelles (Brown clustering et Word2Vec). Ces algorithmes distributionnels sont non supervisés : ils se fondent sur la cooccurrence des mots et de leur voisinage pour regrouper les mots partageant une distribution similaire et créant ainsi des classes sémantiques. Chaque lexique existe en deux versions : création à partir des ensembles *seeds1* et *seeds2*. Dans le cas de Word2Vec, le processus est différent pour les deux séries de codes.

**Lexique.org** est un lexique existant (*http ://lexique.org*) construit par des psycholinguistes. Il contient des liens entre mots morphologiquement liés. Nous utilisons ce lexique pour enrichir l'ensemble initial de seeds avec leurs familles morphologiques. Ainsi *nerveux* est enrichi avec *nerveusement* et *nervosité*. Ce lexique apporte 36 termes.

**Wikidata** (*http ://wikidata.org*) est une base de connaissances sémantiques d'ordre général, dont le rôle est d'héberger et structurer le contenu sémantique de Wikipédia et d'autres projets Wikimédia. Cette base rend ces connaissances sémantiques accessibles à d'autres applications du web sémantique. Cette ressource contient plus de 24 millions d'objets, dont au moins 8 000 objets dans le domaine médical. Plusieurs étapes sont nécessaires pour l'exploitation de cette ressource : (1) Extraction des objets représentant des maladies. Wikidata contient des propriétés telles que *subclass of disease* mais celles-ci ne sont pas utilisées systématiquement. Nous proposons donc d'extraire les objets possédant un code CIM10. Nous nous intéressons aux objets dont le code CIM10 est identique au code d'une seed dans l'ensemble *seeds1*. Les codes *seeds2* correspondant à des classes de maladies et non pas à des maladies, nous nous basons toujours sur les codes *seeds1* pour trouver les objets pertinents. (2) Nous utilisons la propriété *alternative label* pour récolter différentes appellations. Ainsi *agoraphobia* (Q174589) est associée aux labels français *agoraphobe* et *peur sociale*. Nous excluons les labels variant uniquement par l'utilisation de diacritiques {*schizophrénie*, *schizophrènie*, *schizophrenie*},

celles-ci étant neutralisées durant l'analyse des posts, mais nous conservons les variations de flexion {*phobie*, *phobies*} afin de pallier un éventuel défaut de lemmatisation. Nous obtenons 61 termes.

**JeuxdeMots** (Lafourcade, 2007) est une ressource existante issue du crowdsourcing qui propose des relations sémantiques entre deux mots. Chaque relation possède un poids, accordé en fonction de la fréquence à laquelle ces mots sont entrés en association par les participants. Le type de relation peut être indiqué, mais ce n'est pas le cas le plus fréquent. Nous utilisons donc les relations de manière non-typée, ce qui correspond généralement à une relation sémantique faible. Nous utilisons également les relations synonymiques et hyponymiques. Nous conservons les premiers 80 % des termes associés à chaque seed, à l'exception de 5 seeds pour lesquelles les mots associés sont particulièrement ambigus (*anxiété, tristesse, angoisse, panique et fatigue*). Pour ces 5 mots, nous conservons uniquement les premiers 30 %. La ressource apporte 1 566 termes.

**Brown clustering** (Brown *et al.*, 1992; Liang, 2005) est un algorithme distributionnel qui permet de générer de nouvelles ressources à partir de notre corpus. Nous avons empiriquement choisi de générer 500 clusters. Chaque terme du corpus est placé de façon ordonnée dans un cluster. Nous avons d'abord soumis à l'algorithme l'ensemble de discussions du forum médicament de Doctissimo, contenant 61% de messages dans la catégorie contraception, et seulement 10% dans la catégorie antidépresseurs et anxyolitiques. En conséquence, les termes liés aux antidépresseurs se retrouvaient regroupés dans un petit nombre de clusters, sans distinction entre les différentes maladies. Afin d'obtenir des distinctions plus fines, nous avons choisi d'utiliser uniquement le corpus des antidépresseurs et anxyolitiques. Nous générons ensuite deux ressources distinctes : *Brown1* est construit en conservant les 30 premiers termes de chaque cluster contenant une seed. *Brown2* est construit de la même façon, mais en ajoutant également les clusters où se trouve un mot de la même famille morphologique qu'une seed, tel que fourni par *Lexique.org*. Le corpus d'évaluation est exclu du corpus d'entraînement de Brown. Les deux ressources résultantes apportent respectivement 397 et 725 termes.

**Word2Vec** (Mikolov *et al.*, 2013a,b) est également un algorithme distributionnel qui permet de créer des clusters autour d'une seed donnée. Afin de pouvoir comparer les résultats avec ceux de Brown, nous avons utilisé le même corpus constitué uniquement de la catégorie antidépresseurs et anxyolitiques de Doctissimo. Cependant, ce corpus étant composé d'environ 800 000 mots seulement, cela peut limiter les performances de word2vec. Le corpus d'évaluation est exclu de ce corpus d'entraînement. Nous utilisons l'algorithme *cbow* avec une fenêtre de 10 mots et l'utilisation de bigrammes. Pour créer un cluster, nous soumettons à l'algorithme une requête pouvant être constituée d'un ou de plusieurs mots. Pour l'ensemble *seeds1*, chaque requête est constituée d'une seule seed. Pour l'ensemble *seeds2*, nous avons combiné toutes les seeds correspondant à un même code en une seule requête. Pour chaque cluster, nous gardons les 30 premiers mots retournés. Nous créons ainsi quatre lexiques :

1. *W2Vseeds*, basé uniquement sur la liste de seeds.
2. *W2Vmorph*, basé sur les seeds ainsi que sur leurs familles morphologiques.
3. *W2Vortho*, où chaque seed est associée à un mot mal orthographié rencontré souvent dans le corpus (nous avons choisi *medcin*) pour former une requête de deux mots.
4. *W2Vcomb* est l'association de *W2Vmorph* et *W2Vortho* : les seeds et leurs familles morphologiques sont utilisées, et chaque terme est également associé à un mot mal orthographié.

Ces lexiques comportent respectivement 313, 616, 242 et 395 termes pour les codes *seeds1* ; 213, 242, 200 et 219 termes pour les codes *seeds2*. La variation de taille entre chaque lexique s'explique par le recoupement entre chaque cluster généré, qui peut être plus ou moins important. Les lexiques générés pour les codes *seeds2* sont plus petits, car moins de requêtes ont été réalisées pour leur création.

La **Combinaison** de toutes ces ressources génère deux lexiques supplémentaires : *Total* correspond à l'addition de tous les lexiques, et *Vote* contient chaque terme apparaissant dans au moins deux ressources différentes. Dans *Vote*, les seeds sont toujours conservées. Avec l'ensemble *seeds1*, ces lexiques contiennent 2 969 et 135 entrées, alors qu'avec l'ensemble *seeds2*, ces lexiques contiennent 2 508 et 130 termes, respectivement.

### 3.3 Rechercher les posts pertinents

En utilisant les seeds décrites à la section 2.2 et les ressources sémantiques construites précédemment, nous recherchons les messages de forum contenant une mention des maladies étudiées. Nous effectuons cette recherche en enrichissant la liste de seeds avec les termes collectés dans les ressources sémantiques. Les termes récoltés sont rassemblés dans des lexiques. Les termes de ces lexiques sont associés avec le code CIM10 de la seed qui a permis de les collecter. Ce code peut être précisément le code de la maladie (*seeds1*) ou un code plus générique, désignant plusieurs maladies proches (*seeds2*). Nous évaluons ensuite la qualité de l'étiquetage selon deux niveaux de granularité : au niveau des phrases et au niveau des posts, le niveau des phrases étant supposé être plus précis, mais aussi plus difficile à traiter.

## 4 Résultats et discussion

### 4.1 Acquisition des ressources sémantiques

TABLE 1 – Desciption des ressources sémantiques obtenues

| *Lexique* | *Taille seeds1* | *Taille seeds2* | *Exemples* |
|---|---|---|---|
| *seeds* | 23 | 23 | *crise d'angoisse* |
| *Lexique.org* | 36 | 36 | *angoissant* , *angoissé* |
| *Wikidata* | 61 | 61 | *attaque de panique* |
| *JdM* | 1 566 | 1 566 | *convulsion* , *crampe* , *médicament* |
| *Brown1* | 397 | 397 | *dépersonalisation* , *hystérie* , *alzheimer* |
| *Brown2* | 725 | 725 | *hystérique* , *stresser* , *suicidaire* , *bailler* |
| *W2V seeds* | 313 | 213 | *spasmophilie* , *violent* , *gros* |
| *W2V morph* | 616 | 242 | *cercle vicieux* , *trembler* , *devenir fou* |
| *W2V ortho* | 258 | 200 | *spasmo* , *dangoisse* , *fesais* |
| *W2V comb* | 395 | 219 | combinaison de *W2Vmorph* et *W2Vortho* |
| *Total* | 2 969 | 2 508 | fusion de tous les lexiques |
| *Vote* | 135 | 130 | vote de tous les lexiques |

Dans la table 1, nous décrivons les lexiques générés à partir des ressources existantes et du corpus. Nous indiquons leur taille et donnons quelques exemples de termes dans la dernière colonne. Les exemples sont donnés pour la seed *panique*, *crise d'angoisse*, dont le code CIM10 est F41.0. Nous avons tenté de sélectionner des exemples représentatifs du contenu de chaque ressource. Nous constatons que les termes proposés par les différentes méthodes sont plus ou moins sémantiquement

proches de la seed. *Lexique.org*, qui propose des mots de la même famille morphologique, contient des mots très proches. *Wikidata* contient également des mots très proches. Les autres ressources peuvent contenir des mots plus distants ( *convulsion, crampe, médicament, alzheimer, bailler, spasmophilie, violent, gros, trembler, fesais...*) aux côtés de mots plus pertinents (*dépersonalisation, hystérie, hystérique, stresser, suicidaire, cercle vicieux, devenir fou, spasmo, dangoisse...*).

Le lexique issu de *Wikidata* contient 61 termes, qui peuvent être des expressions composées de plusieurs mots : *dépression, dépression clinique, dépression caractérisée, dépression majeure*. Nous trouvons également des variantes morphologiques et syntaxiques intéressantes (*intolérance au glucose, intolérance aux glucides, intolérance glucidique*) ainsi que des fautes d'orthographe. Toutes ces variantes ne sont pas forcément présentes dans notre corpus.

Le lexique *JeuxdeMots* contient 1 566 termes. Ces termes peuvent être plus ou moins sémantiquement proches des seeds. Par exemple pour la seed *dépression* nous obtenons :

— *anxiété, tristesse, angoisse, panique, fatigue*
— *pauvreté, rater, malheureusement, sanglot, résigner, souffrira*

Si la première série contient des termes pertinents et intéressants, la deuxième série risque d'apporter surtout du bruit.

Les lexiques *Brown* tendent à regrouper les termes d'une même classe sémantique : les maladies, les médicaments, etc. Par exemple, *anxiété* se trouve dans le même cluster que *hypocondrie, timidité, hyperphagie, isolement, urticaire...* . Nous présumons que ces clusters assez larges nous seront moins utiles pour notre tâche.

L'algorithme *Word2Vec* donne généralement des candidats assez proches. Pour l'exemple de *dépression*, les premiers termes obtenus sont *dépressif* et *anxieux*. Une autre piste qui sera explorée dans des travaux futurs est de réutiliser les premiers termes pour obtenir davantage de candidats pertinents, en procédant de proche en proche. Ainsi, *anxieux* nous donne à son tour *nerveux, anxiété, stresser, panique, déprime*. En ajoutant aux seeds leur famille morphologique, nous avons construit les lexiques *W2V morph* et *W2V comb*. Construire une requête constituée de mots de classes morphologiques différentes nous permet de neutraliser la tendance de l'algorithme à sélectionner des mots de même classe morphologique que la requête. Nous avons utilisé des mots mal orthographiés pour générer *W2V ortho* et *W2V comb* afin d'obtenir des mots également mal orthographiés dans le cluster, comme exemplifié dans la table 1.

Le lexique *Total* contient bien sûr le plus grand nombre d'entrées, indiquant que nos lexiques sont complémentaires et peuvent chacun apporter de nouveaux termes. Mais nos lexiques se chevauchent également. Voici le nombre d'entrées exclusives à chaque lexique : 47 dans *Wikidata*, 1 472 dans *JdM*, 625 dans les deux *Brown* regroupés et 639 dans les quatre *W2V* regroupés.

## 4.2 Formes recherchées

Le trouble le plus représenté dans le corpus d'évaluation est la dépression, qui constitue un bon exemple de la variété des formes recherchées. 103 extractions dans 96 phrases ont été identifiées dans le corpus d'évaluation. Les termes *dépression, anti-dépresseur* et *dépressif*, ainsi que leurs variations orthographiques et abréviations, constituent 76% des termes à identifier. Avec une simple neutralisation de la casse, sans autre prétraitement, ces trois termes totalisent 40 formes distinctes. Le terme *anti-dépresseur* en particulier totalise 13 formes différentes, à cause de variations orthographiques pouvant se combiner les unes avec les autres : présence ou non de l'accent sur *dépression*, forme au

TABLE 2 – Résultats pour l'ensemble *seeds1*

| | Message | | | | Phrase | | | |
|---|---|---|---|---|---|---|---|---|
| *Lexique* | *TP* | *P* | *R* | *F* | *TP* | *P* | *R* | *F* |
| *seeds* | 81 | 0.920 | 0.529 | 0.672 | 102 | 0.918 | 0.430 | 0.586 |
| *Lexique.org* | 94 | **0.921** | 0.614 | **0.737** | 123 | **0.924** | 0.518 | **0.664** |
| *Wikidata* | 86 | 0.914 | 0.562 | 0.696 | 108 | 0.915 | 0.455 | 0.608 |
| *Jdm* | 115 | 0.392 | **0.751** | 0.515 | 135 | 0.242 | **0.569** | 0.340 |
| *Brown1* | 91 | 0.590 | 0.594 | 0.592 | 108 | 0.486 | 0.455 | 0.470 |
| *Brown2* | 91 | 0.590 | 0.594 | 0.592 | 108 | 0.486 | 0.455 | 0.470 |
| *w2v seeds* | 82 | 0.577 | 0.535 | 0.555 | 102 | 0.528 | 0.430 | 0.474 |
| *w2v morph* | 91 | 0.476 | 0.594 | 0.529 | 108 | 0.421 | 0.455 | 0.438 |
| *w2v ortho* | 85 | 0.658 | 0.555 | 0.602 | 105 | 0.668 | 0.443 | 0.532 |
| *w2v comb* | 84 | 0.677 | 0.549 | 0.606 | 104 | 0.688 | 0.438 | 0.536 |
| *vote* | 109 | 0.698 | 0.712 | 0.705 | 135 | 0.613 | 0.569 | 0.590 |
| *total* | 135 | 0.289 | 0.882 | 0.436 | 165 | 0.170 | 0.696 | 0.273 |

pluriel ou au singulier, séparateur entre *anti* et *dépresseur*, qui peut être un espace, un tiret ou être absent. Il existe aussi l'abréviation *AD*, qui peut être mise au pluriel *ADs*, ainsi que les abréviations *anti dep* et *anti dd*, chacune trouvée une seule fois.

Neutraliser les variations de la liste suivante permet de diminuer le nombre de formes distinctes pour *dépression, antidépresseur* et *dépressif* de 40 formes à 26 formes.

- — diacritique (dépression / depression)
- — flexion (dépressif / dépressive)
- — séparateur dans un mot composé (antidépresseur / anti dépresseur / anti-dépresseur)
- — séparateur entre les mots : à la suite d'une erreur typographique, l'espace entre le terme recherché et le mot voisin est manquant

15% des annotations restantes sont des phrases pour lesquelles il est difficile d'identifier un terme isolé comme synonyme du trouble recherché. On trouve par exemple *"moins de joie de vivre qu'avant"* et *"je n'arrive plus a reflechir ni imaginer"*.

Les 10% restants sont des termes plus variés, qui sont présents une à trois fois dans le corpus d'évaluation. On trouve par exemple *déprime* (3 occurences), *idées noires* (3 occurences) et *mal-être* (2 occurences).

Sur l'exemple de *dépression*, il apparaît que la majorité des termes à retrouver sont des variations orthographiques de seulement 3 termes différents. Sur ces trois termes, *dépression* est la seed, *dépressif* fait partie de la même famille morpho-syntaxique, et *anti-dépresseur* est formé sur la même base, mais n'est pas un synonyme de la seed.

## 4.3 Recherche des messages pertinents et évaluation

Dans les tables 3 et 2, nous présentons les résultats de la recherche de messages pertinents pour les maladies étudiées. Pour chaque lexique et chaque unité (phrase ou message) nous indiquons les mesures suivantes : nombre de vrais positifs TP, précision P, rappel R et F-mesure F (Sebastiani,

TABLE 3 – Résultats pour l'ensemble *seeds2*

| | Message | | | | Phrase | | | |
|---|---|---|---|---|---|---|---|---|
| *Lexique* | *TP* | *P* | *R* | *F* | *TP* | *P* | *R* | *F* |
| *seeds* | 82 | 0.911 | 0.522 | 0.663 | 98 | 0.859 | 0.388 | 0.535 |
| *Lexique.org* | 93 | 0.902 | 0.592 | **0.715** | 120 | **0.869** | 0.476 | **0.615** |
| *Wikidata* | 83 | **0.912** | 0.528 | 0.669 | 100 | 0.862 | 0.396 | 0.543 |
| *JDM* | 129 | 0.452 | **0.821** | 0.583 | 159 | 0.296 | **0.630** | 0.403 |
| *Brown1* | 93 | 0.654 | 0.592 | 0.622 | 105 | 0.541 | 0.416 | 0.470 |
| *Brown2* | 103 | 0.490 | 0.656 | 0.561 | 125 | 0.378 | 0.496 | 0.429 |
| *w2v seeds* | 94 | 0.556 | 0.598 | 0.576 | 109 | 0.480 | 0.432 | 0.455 |
| *w2v morph* | 100 | 0.628 | 0.636 | 0.632 | 109 | 0.542 | 0.432 | 0.481 |
| *w2v ortho* | 85 | 0.720 | 0.541 | 0.618 | 102 | 0.698 | 0.404 | 0.512 |
| *w2v comb* | 86 | 0.722 | 0.547 | 0.623 | 100 | 0.689 | 0.396 | 0.503 |
| *vote.txt* | 102 | 0.75 | 0.649 | 0.696 | 130 | 0.702 | 0.515 | 0.594 |
| *total.txt* | 136 | 0.349 | 0.866 | 0.498 | 171 | 0.222 | 0.678 | 0.335 |

2002). Le meilleur résultat pour chaque mesure est marqué en gras. Comme seulement 100 messages ont pu être annotés pour l'évaluation, la représentativité de ces résultats est limitée. Aucune validation croisée n'a été effectuée.

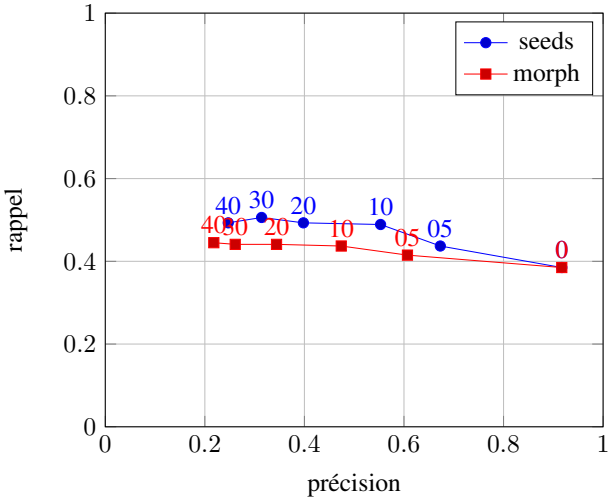
Comme attendu, identifier les messages est plus facile qu'identifier les phrases : la granularité est alors moins fine et le nombre de mots potentiellement pertinents plus important. De même, la granularité moins fine au niveau des seeds rend les résultats pour *seeds2* légèrement meilleurs que pour *seeds1*.

Nous notons une nette différence entre *seeds*, *Lexique.org* et *Wikidata* d'une part, et *JDM* et les méthodes distributionnelles d'autre part, en ce qui concerne la précision. Ce n'est pas surprenant : comme les premières ressources restent confinées aux synonymes et quasi-synonymes, la précision est donc excellente.

Pour *JDM* et les ressources distributionnelles, les termes obtenus étant classés par poids, il est possible d'ajuster les résultats en modifiant le nombre de termes sélectionnés pour chaque seed. La figure 1 montre l'évolution du rappel et de la précision selon le nombre de termes par seed sélectionné. Les paramètres suivant ont été utilisés : le lexique *W2V morph*, l'ensemble *seeds1*, les résultats calculés au niveau des phrases.

Dans le graphe de la figure 1, prendre "0 terme" correspond à ne prendre que les seeds. Nous observons que la précision diminue rapidement pour une faible amélioration du rappel. Au regard de ces résultats, nous avons choisi de conserver les 30 premiers termes pour nos résultats, car cette configuration est la plus représentative de la ressource, malgré sa faible précision. En effet, nos ressources fournissent des termes pertinents, améliorant le rappel, mais ils sont mêlés à des termes bien trop génériques, ce qui déteriore la précision assez rapidement. Par exemple, parmi les termes associés à *dépression* par Word2Vec, nous trouvons *trouble* en troisième position, alors que les premiers termes associés à *angoisse* sont, dans l'ordre, *sentiment, crise, crainte*. A cause de ces termes génériques haut placés, même si nous en sélectionnons un très petit nombre, cela ne permettrait pas de conserver une précision satisfaisante. Il est donc nécessaire d'utiliser une méthode complémentaire pour filtrer les termes fournis par ces ressources. Plusieurs sont proposées par (Claveau *et al.*, 2014) et seront explorées dans de futurs travaux.

FIGURE 1 – Résultats obtenus avec *W2V* selon le nombre de termes retenus par seed




Le meilleur rappel est obtenu avec le lexique *JDM*, qui est le plus volumineux. *Brown2* et *W2V morph*, tous deux basés sur l'exploitation de seeds et de leurs familles morphologiques, sont le plus souvent second et troisième.

Si nous observons les termes qui apportent le plus de faux négatifs dans le corpus d'évaluation, nous remarquons que le trouble le plus fréquent est la dépression, avec 46% des faux négatifs pour Lexique.org (seed1, phrase). C'est aussi le trouble que l'annotateur a généralement associé aux expressions les plus vagues, à cause de sa généricité. On trouve notamment des expressions telles que *"un peu moins de joie de vivre qu'avant","baisse de moral"* ou *"je ne me reconnais plus, plus rien de m'intéresse, je n'arrive plus à réfléchir ni à imaginer"*.

Parmi les faux négatifs de dépression, 35%, soit 19 annotations, correspondent au terme *anti-dépresseur* ou à une variante de celui-ci (*AD*, *anti dep*...). En effet, dans de nombreux messages, une personne demande conseil sur un médicament sans citer la maladie pour laquelle le médicament serait pris. On a par exemple un message dont le titre est le nom d'un médicament, et le corps de texte uniquement *"Qui a eu une amélioration avec cet AD ?"*. Dans ce cas, il existe un lien implicite entre le médicament et la maladie, et l'annotation est donc correcte. Mais *antidépresseur* n'étant pas un synonyme de *dépression* et n'étant pas utilisé dans le même contexte morpho-syntaxique, ce terme n'est retrouvé que dans deux de nos lexiques : Jeux de Mots et w2v morph. Le cas de w2v morph est le plus intéressant et celui sur lequel nous nous concentrons : parmi les variantes morphologiques utilisées comme seed, on trouve *anxyolitique*, dérivé de *anxiété*. Or, *anxyolitique* est le nom d'une classe de médicaments, auquel word2vec associe facilement *antidépresseur*. *Antidépresseur* est donc présent dans ce lexique, mais incorrectement associé à *anxiété*. Les variantes *AD* et *antidep* sont également présentes, mais incorrectement classifiées. Il apparaît donc que word2vec peut trouver les variantes orthographiques d'*anti-dépresseur* à partir de ce terme, mais pas à partir du terme *dépression*.

Le regroupement de termes au sein de *seeds2* ne produit pas d'amélioration remarquable excepté

pour les lexiques *W2V*, qui ont été générés avec des requêtes multi-mots. Ces lexiques sont alors moins fournis que leurs équivalents *seeds1*. Nous obtenons donc un rappel plus faible, mais largement contrebalancé par l'amélioration de la précision.

*Lexique.org* obtient la meilleure F-mesure dans toutes les configurations, principalement due à son excellente précision. Si nous souhaitons accorder davantage d'importance au rappel, *Vote* fournit un meilleur équilibre entre les deux mesures.

# 5 Conclusion et travaux futurs

Nous avons proposé de construire et enrichir automatiquement des lexiques pour la détection de messages de forums parlant de troubles de l'humeur. La particularité de ce travail est la nécessité d'analyser et traiter des messages écrits par des patients sur un forum de discussion, qui contiennent un vocabulaire et des expressions spécifiques. Nous avons proposé d'utiliser des ressources existantes et de créer de nouvelles ressources grâce aux méthodes non supervisées exploitant notre corpus. La ressource *Lexique.org* fournit la meilleure F-mesure, alors que la ressource *Vote*, qui contient les termes proposés par au moins deux ressources, permet d'obtenir le meilleur équilibre entre précision et rappel tout en gardant une F-mesure élevée.

Les ressources existantes que nous avons exploitées contiennent des quasi-synonymes ou bien de nombreux termes faiblement liés sémantiquement ; tandis que les méthodes distributionnelles fournissent des termes pertinents mais aussi des termes plus généraux qui déteriorent la précision. Pour améliorer les ressources distributionnelles, il est possible d'utiliser des corpus plus grands afin d'améliorer la granularité des clusters. Un filtrage supplémentaire de clusters obtenus est également souhaitable. Cela permettrait d'utiliser davantage de critères de sélection, notamment en ce qui concerne la spécificité des termes. Nous comptons également étudier davantage le fonctionnement de Word2Vec et différentes méthodologies pour la combinaison de termes dans les requêtes adressées à Word2Vec afin d'obtenir des clusters plus précis et plus diversifiés. Enfin, nous étendrons notre méthode à un éventail plus large de maladies afin d'évaluer son adaptabilité dans un contexte plus varié.

# Remerciements

La présente publication s'inscrit dans le programme *Drugs Systematized Assessment in real-liFe Environnement (DRUGS-SAFE)* financé par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé. Cette publication ne représente pas nécessairement l'opinion de l'ANSM.

Je remercie vivement Natalia Grabar et Frantz Thiessard pour leur supervision et leur aide à la rédaction, Pierre Simonetti pour l'annotation, Gayo Diallo, Vianney Jouhet et toute l'équipe ERIAS.

# Références

ADAM C., FABRE C. & MULLER P. (2013). Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *Traitement Automatique des Langues*, **vol.**

**54**(n 1), 71–97.

BROWN P., DESOUZA P., MERCER R., DELLA PIETRA V. & LAI J. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

BUDANITSKY A. & HIRST G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, **32**(1), 13–47.

CHOY Y., CHIU C. & SONTAG D. (2016). Learning low-dimentional representations of medical concepts. In *AMIA Jt Summits Transl Sci Proc*, volume 2016, p. 41–50.

CLAVEAU V., KIJAK E. & FERRET O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In *21ème conférence sur le Traitement Automatique des Langues Naturelles, TALN 2014*, p.Ĩ2, Marseille, France.

EHOLIÉ S., TAPI NZALI M. D., BRINGAY S. & JONQUET C. (2016). MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums. In A. PASCHKE, A. BURGER, A. SPLENDIANI, M. MARSHALL & P. ROMANO, Eds., *SWAT4LS : Semantic Web Applications and Tools for Life Sciences*, Amsterdam, Netherlands.

GAUDUCHEAU N. (2008). La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives. *Bulletin de psychologie*, **496**, 389–404.

GRABAR N. & HAMON T. (2014). Automatic extraction of layman names for technical medical terms. In *2014 IEEE International Conference on Healthcare Informatics*, p. 310–319.

KHARE R., GOOD B. M., LEAMAN R., SU A. I. & LU Z. (2016). Crowdsourcing in biomedicine : challenges and opportunities. *Briefings in Bioinformatics*, **17**(1), 23.

LAFOURCADE M. (2007). Making people play for lexical acquisition. *7th Symposium on Natural Language Processing*.

LIANG P. (2005). *Semi-Supervised Learning for Natural Language*. Master, Massachusetts Institute of Technology, Boston, USA.

LIU F., CHEN J., JAGANNATHA A. & YU H. (2016). Learning for biomedical information extraction : Methodological review of recent advances. *CoRR*, **abs/1606.07993**.

MESSAI R., ZENG Q., MOUSSEAU M. & SIMONET M. (2006). Building a bilingual french-english patient-oriented terminology for breast cancer. In *MedNet*, Toronto, Canada.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

MIKOLOV T., SUSTKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*.

OMS (1995). *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*. Organisation mondiale de la Santé, Genève.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, p. 44–49, Manchester, UK.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

ZENG Q. & TSE T. (2006). Exploring and developing consumer health vocabularies. *JAMIA*, **13**, 24–29.

# Création automatique d'une grammaire syntaxico-sémantique

Émilie Colin[1]
(1) équipe SYNALP, Université de Lorraine/LORIA, Nancy, France
emilie.colin@loria.fr

Résumé

Nous proposons une nouvelle méthode pour la création automatique de grammaires lexicalisées syntaxico-sémantiques. A l'heure actuelle, la création de grammaire résulte soit d'un travail manuel soit d'un traitement automatisé de corpus arboré. Notre proposition est d'extraire à partir de données VerbNet une grammaire noyau (formes canoniques des verbes et des groupes nominaux) de l'anglais intégrant une sémantique VerbNet. Notre objectif est de profiter des larges ressources existantes pour produire un système de génération de texte symbolique de qualité en domaine restreint.

Abstract

**Syntactic-semantic grammar automatic creation**

We propose a new method to automaticaly create a syntactico-semantic lexicalized grammar. For the moment, this work is done either manually or with the support of an automated processing of annotated corpus. Our proposal is to extract verbal, nominal and auxiliary trees from VerbNet. Our goal is to take advantage of the vast resources available to produce a system of quality text symbolic generation in a restricted domain.

Mots-clés : grammaire, génération, extraction, ressources.

Keywords: grammar, generation, extraction, resources.

## 1 Introduction

Une grammaire syntaxico-sémantique décrit la relation entre langue naturelle, syntaxe et sémantique. Couplée avec un algorithme de réalisation de surface, elle peut notamment être utilisée pour générer des phrases (Narayan & Gardent, 2012; Cahill & Van Genabith, 2006; Carroll & Oepen, 2005). Pour une représentation sémantique donnée, l'algorithme de réalisation de surface produira la ou les phrases associée(s) par la grammaire à cette représentation.

Les méthodes existantes pour la création de telles grammaires sont souvent manuelles (Gardent, 2006) ou spécifiques à un domaine donné (Wong & Mooney, 2007; Gyawali, 2016). Dans cet article, nous proposons une méthode qui permet de créer automatiquement une grammaire noyau de l'anglais avec une sémantique générique intégrant les rôles thématiques de Verbnet (Kipper *et al.*, 2008). La grammaire est une grammaire d'arbres adjoints lexicalisée à traits (FB-LTAG, (Vijay-Shanker & Joshi, 1988)) et est créée à partir des exemples contenus

dans VerbNet et de leur annotation syntaxico-sémantique. Afin de faciliter la maintenance et l'extension de cette grammaire (e.g., extension à des formes non canoniques comme les propositions relatives), l'extraction automatique produit une méta-grammaire (Candito, 1998, 1996) plutôt qu'une grammaire. Les fragments d'arbres et de sémantiques produits par l'extraction sont ensuite assemblés par un compilateur de métagrammaires existant (XMG, (Crabbé *et al.*, 2013), section 4) pour créer les arbres de la grammaire FB-LTAG.

Une méthode pour créer non pas manuellement mais automatiquement une grammaire généraliste est un objectif nouveau. La création manuelle implique un travail fastidieux et une faible adaptabilité, cette dernière étant le talon d'Achille des grammaires spécifiques à un domaine. La grammaire générée à partir de *treebanks* est d'un autre ressort. Elle offre une description d'un corpus. La technique que nous proposons permet l'extraction d'une grammaire adaptable, ciblant la possibilité de générer des phrases simples, à partir des données (exemples et annotations syntaxico-sémantique) de Verbnet (section 6). Nous montrons que cette grammaire permet de (re)-générer les exemples contenus dans Verbnet avec un score BLEU variant de 0.93 à 0.99 suivant la longueur des phrases (section 7).

## 2 Travaux connexes

Pour ce qui touche à la génération de grammaire noyau, la recherche oscille entre savoirs experts et restriction de l'espace syntaxico-sémantique. La création de grammaire exploitable en traitement automatique des langues est loin d'être un objectif neuf. Pour l'heure, il existe différentes méthodes : manuelle, en s'appuyant sur le travail de linguistes, automatique, en restreignant la tâche par le domaine (usage de données métier/d'un corpus annoté).
Le coût des savoirs experts reste notable, même dans un cadre où l'assistance par ordinateur est maximisée (Xia, 2001) : tout doit être pré-pensé, structuré "manuellement". Il en découle que la révision de la grammaire est difficile.
Se centrer sur un domaine particulier permet d'adapter aux données un algorithme d'étiquetage sémantique (Wong & Mooney, 2007) ou de développer des statistiques d'usage cadrées par l'expressivité et le vocabulaire propres au domaine (Gyawali, 2016).
Ciblant deux buts relativement proches, une autre méthode doit être mentionnée, une méthode stochastique. Celle-ci ne cible pas la création de grammaire noyau. Il s'agit de l'exploitation de *treebanks* : de corpus arborés (Chiang (2000); Xia (2001)). Cela permet la création d'une grammaire relativement couvrante, du moins vis-à-vis du corpus de référence. Par contre, nous dit Xia (2001), la consistance n'est pas garantie et la flexibilité reste pauvre. Pour un verbe utilisé d'une manière donnée, on en trouve douze répondant aux mêmes règles [douze est le nombre de verbes moyen par classe obtenu par Kipper *et al.* (2008)] : un corpus ne permet pas d'apprendre cela.

## 3 La grammaire, formalisme utilisé : FB-LTAG

Pour décrire la grammaire de l'anglais ou d'une autre langue de façon exploitable par un ordinateur, il est nécessaire d'utiliser un formalisme syntaxique précis et adéquat.

L'analyse syntaxique en TAG [1] se réalise en temps polynomial et permet notamment de décrire des contraintes entre constituants éloignés dans la phrase. L'expressivité est grande (classe de langage légèrement sensible au contexte). Enfin, il existe des outils pour la génération, comme GenI (Gardent & Kow, 2007a).

FB-LTAG [2] est une variante de TAG où chaque arbre TAG est associé avec un mot et décoré avec des structures de traits. Gardent & Kallmeyer (2003) montrent comment cette variante permet d'approcher la sémantique de manière compositionnelle en conjuguant sémantique et grammaire. C'est le formalisme que nous avons choisi.

Les arbres de la grammaire sont exploitables par le biais d'une sémantique. La phrase (1a) a pour sémantique plate la sémantique (1b).

(1) a. *A fire rage in the mountains.*
b. a(`w`) fire(`w`) rage(`e1`) Theme(`e1 w`) in(`e1 j`) the(`j`) mountains(`j`)

TABLE 1 – Sémantique plate

La phrase 1a utilise l'évènement `e1` *rage*. *Rage* est un évènement défini dans un lexique (cf figure 2) comme attendant un thème et comme pouvant s'ancrer dans un arbre où il sera spécifiable par une préposition. *In* est défini dans le lexique comme préposition pouvant mettre en relation $x$ avec un évènement. Cet $x$ ne pourra être que un groupe nominal (NP) car le groupe prépositionnel de l'arbre exploitable par *rage* a une définition stricte (voir Figure 1).

```
*ENTRY: rage
*CAT: v
*SEM: unaryRel[rel=rage,theta1=Theme]
*FAM: n0VPrepPN
*EX: {A_fire_raged_in_the_mountains.}
```

TABLE 2 – extrait lexical minimaliste : *rage*

Nous n'avons pas utilisé FB-LTAG directement. Nous l'avons fait générer par un compilateur de méta-grammaire, XMG, que nous allons présenter.

# 4 La méta-grammaire, XMG

Les méta-grammaires sont nées de la volonté d'offrir un langage simple d'utilisation, de pouvoir directement décrire ces notions syntaxiques de haut-niveau que sont les notions de sujet, d'objets, de verbe.

XMG est un terme faisant référence à la fois à un langage formel et à un compilateur éponyme permettant la génération de grammaires à partir de leur description (ou *métagrammaire*). Une définition formelle (syntaxe et sémantique) du langage XMG est donnée dans (Crabbé

1. TAG : Tree Adjoining Grammar, grammaire d'arbres adjoint
2. FB-LTAG : Feature-Based Lexicalized TAG, grammaire d'arbres adjoints lexicalisée basée sur des propriétés sémantiques

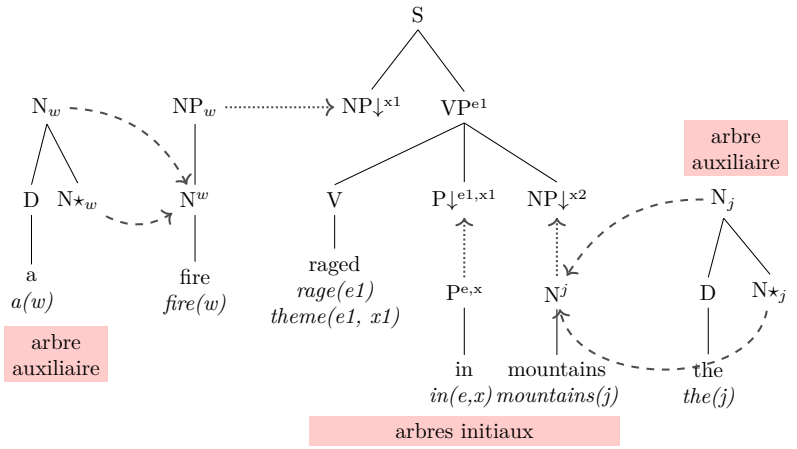



FIGURE 1 – "A fire raged in the mountains"

*et al.*, 2013). Nous résumons ici brièvement les caractéristiques de ce formalisme utilisés dans le cadre de notre travail.

XMG permet de spécifier et de combiner des classes qui intuitivement, décrivent des fragments d'arbres, des fragments de représentations sémantiques et le partage de variables entre arbres et représentations sémantiques. Les fragments sont représentés sous forme de formules de logique de description d'arbres (Rogers & Vijay-Shanker, 1994). Les classes peuvent être combinées par héritage, conjonction et disjonction. Un système de couleur est également utilisable qui permet de contraindre l'unification de variables de nœuds d'arbre : un nœud rouge ne peut pas être identifié avec un autre nœud, un nœud noir peut être identifié avec un ou plusieurs nœuds blancs, et un nœud blanc doit être identifié avec un nœud noir. La Figure 2 illustre ces mécanismes avec de haut en bas, des fragments d'arbres, des fragments de représentations sémantiques et les équations permettant d'identifier des variables apparaissant dans les arbres d'une part et dans la sémantique d'autre part.

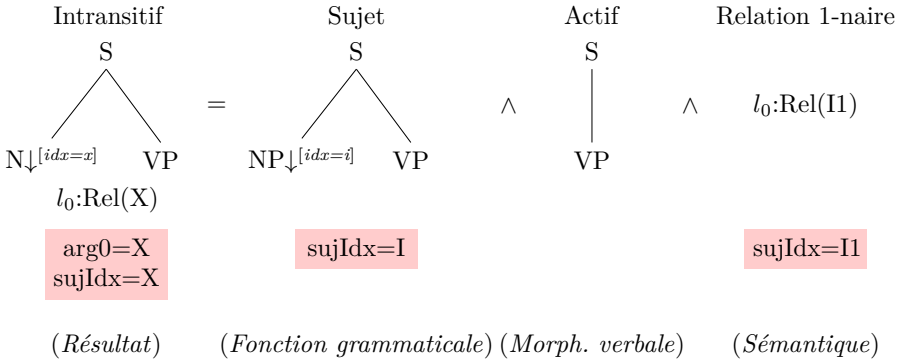



FIGURE 2 – Fragments XMG et leur combinaison

La description XMG d'une grammaire est souvent appelée une métagrammaire.

L'utilisation de XMG visait deux buts : réduire le coût de mise en œuvre de notre méthode,

par le biais de production intermédiaires intelligibles, et offrir un résultat aisé à maintenir : les fichiers XMG non compilés sont faciles à modifier.

XMG nous permet de gérer les informations sémantiques : la classe `unaryRel` correspond à la sémantique plate d'une relation unaire et permet le rapprochement des éléments lexicaux avec le contenu admis par des nœuds. Sa définition permet à un générateur le couplage de la sémantique avec les arbres compilés.

XMG nous a permis de décrire dynamiquement des relations unaires ou binaires, qu'un rôle attendant selon verbnet un objet de type '*adj*' pouvait accueillir aussi bien un participe passé (*vbd*) qu'un adjectif (*jj*), mais aussi qu'un adjectif pouvait modifier un nom pour devenir un groupe nominal, ce qui, après compilation, donne lieu à la création d'un arbre auxiliaire.

Nous avons utilisés les données VerbNet pour construire des classes XMG. Les notions de *modifiers* (un adjectif vient modifier le sens d'un nom) ont pu être gérées semi-automatiquement grâce à l'expressivité d'XMG, en partant du découpage des groupes syntaxiques puisés dans les phrases-exemple VerbNet et isolés informatiquement. La factorisation complètement automatisée serait, à ce niveau, trop complexe (erreurs d'étiquetage des composants de groupes syntaxiques). Faute de place, nous nous étendrons peu sur ce sujet ici, mais nous tenons à dire que la possibilité de décrire des alternatives entre fragments se prêtent fort bien à l'automatisation. La possibilité de nommer, d'architecturer les classes facilitent retours et adaptations.

# 5 VerbNet

Kipper *et al.* (2000) ont créé un lexique verbal exploitable pour les grammaires d'arbres adjoints. VerbNet 3.2, celui que nous utilisons, décrit 3695 verbes au sein d'un système de classes. Kipper *et al.* ont utilisé TAG pour construire leur lexique hiérarchisé au sein de classes.

Le système de classes de VerbNet est la structure accueillant les descriptions syntaxico-sémantique des verbes traités. Toute classe peut avoir 0 à $n$ sous-classes. Une sous-classe est une classe qui hérite d'une autre classe.
Une classe ou sous-classe comprend un certain nombre de descriptions syntaxico-sémantiques. Nous allons voir comment ces descriptions sont structurées à l'aide d'identifiants syntaxiques et sémantiques, chaque identifiant étant doté de restrictions si nécessaire.

Chaque identifiant syntaxique (NP ou V, par exemple) est associé à un identifiant sémantique (tel NP est l'*Agent* dans la phrase, tel autre le *Thème*). Chaque identifiant, selon sa catégorie, peut être associé à des restrictions syntaxiques (ex : pluriel obligatoire) ou sémantiques (l'agent est *animate* ou *organization*).

Il y a 274 classes-racine dans VerbNet 3.2. Un verbe peut appartenir et en règle générale, appartient à plusieurs classes. La Figure 3 illustre les relations d'héritage de l'une des classes, *entity_specific_modes_being-47.2*.

Un verbe (tel *rage*) défini dans la classe *entity_specific_modes_being-47.2* a pour structures valides celles apportées par sa classe propre (*structure_a, structure_b, ...*). D'autres verbes

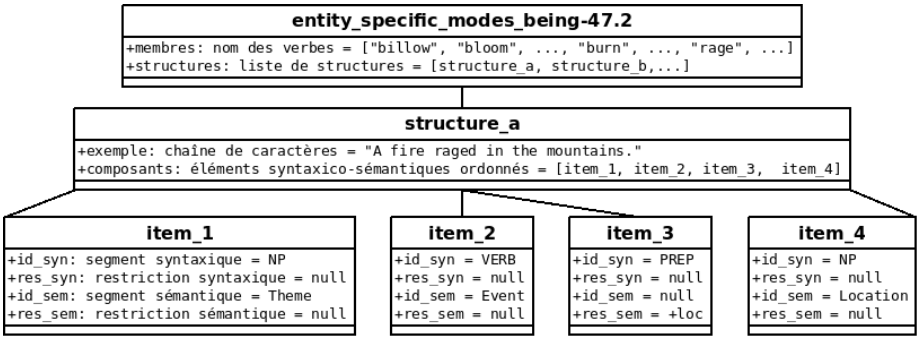



FIGURE 3 – illustration de la classe *entity_specific_modes_being-47.2*

(ex : *burn*) appartiennent à cette classe. Les structures décrites pour la classe sont aussi valides pour eux. Ces verbes peuvent être définis dans des sous-classes. Cela se produit quand un ou plusieurs verbes peuvent utiliser les règles de constructions de la classe-mère mais que d'autres structures spécifiquement exploitables par eux ont été référencées.
Un verbe (tel *report*) est défini dans la sous-classe *characterize-29.2-1-2* (en bas à droite Figure 4). Il a pour structures valides celles apportées par sa classe propre, mais il a aussi pour structure valides celles déclarées dans les classes dont il hérite : *characterize-29.2-1* et *characterize-29.2*. En bas à gauche Figure 4, *characterize-29.2-1-1*, qui existe en parallèle et ne le concerne pas, est également héritière de *characterize-29.2-1* et *characterize-29.2*.

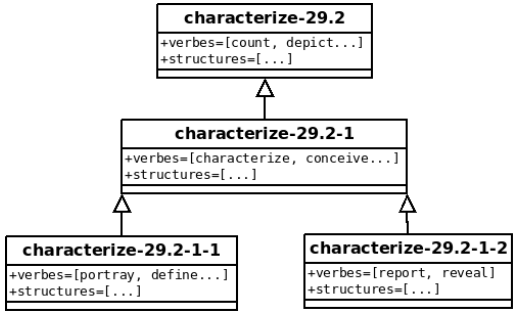



FIGURE 4 – Représentation de l'héritage sur *characterize-29.2*

L'extraction FB-LTAG, expliquée dans la partie suivante, est facilitée par cette organisation des données.

# 6 Extraction depuis VerbNet

Cette partie est consacrée à l'extraction automatique de la grammaire et du lexique verbal.

## 6.1 Vue d'ensemble

Nous avons vu que VerbNet décrit pour chaque classe les structures syntaxico-sémantiques utilisables par un nombre fini de verbes. VerbNet illustre chaque structure syntaxico-sémantique par un exemple (cf Figure 3).

Une structure syntaxico-sémantique est un quintuplet composé de :
- deux séries parallèles de segments syntaxiques,
- deux séries parallèles de segments sémantiques,
- un exemple.

La classe *entity_specific_modes_being-47.2* décrit la structure syntaxico-sémantique montrée dans la Table 3. Les séries de segments syntaxiques sont illustrés par les lignes a, les séries de segments sémantiques par les lignes b. Ces séries sont exemplifiées par la phrase c.

| | | | | références Figure 3 | | | |
|---|---|---|---|---|---|---|---|
| | | | | item_a | item_b | item_c | item_d |
| a | segments syntaxiques | $a^1$ | identifiants | NP | Verb | PREP | NP |
| | | $a^2$ | restrictions | $\emptyset$ | $\emptyset$ | $\emptyset$ | $\emptyset$ |
| b | segments sémantiques | $b^1$ | identifiants | Theme | Event | Prep | Location |
| | | $b^2$ | restrictions | $\emptyset$ | $\emptyset$ | *+loc* | $\emptyset$ |
| c | exemple | | | A fire raged in the mountains. | | | |

TABLE 3 – Structure syntaxico-sémantique adaptée à *A fire raged in the mountains.*

Toutes les informations syntaxiques nous ont été utiles, à la fois pour traiter les phrases exemple (segmentation en fonction des groupes attendus, caractérisation), à la fois pour introduire de fines et nécessaires contraintes syntaxiques. Les restrictions syntaxiques (il y en a 41 différentes) permettent de poser, par exemple, qu'un groupe nominal est systématiquement introduit par *how* pour une structure donnée. Manquant de place pour détailler les restrictions, nous renvoyons le lecteur à Kipper *et al.* (2008) pour plus d'informations.
Les restrictions sémantiques (ex : ligne $b^2$, Table 3) n'ont pas été exploitées. Elles seraient utiles, pour spécifier par exemple lexicalement les adverbes et restreindre les possibles en génération. Ce choix n'est lié qu'à des contraintes de temps. Les identifiants sémantiques nous ont permis de déterminer les rôles sémantiques des arguments syntaxiques et des modifieurs (sujet, COD, compléments circonstanciels...).

Les composants, segments syntaxico-sémantiques, parmi lesquels les items *item_a*, *_b*, *_c* et *_d* introduits Figure 3, sont au nombre de 192. Avec leur aide, nous extrayons de l'ensemble des classes fournies par VerbNet une grammaire d'arbres adjoints noyau reflétant les structures verbales décrites dans cette ressource. La procédure d'extraction se fait selon les étapes suivantes :

- segmentation syntaxico-lexicale : chaque exemple est segmenté en fragments ; le couplage des informations initiales (lignes a, b et c, Table 4) à ces fragments (d) est effectué en phase avec les informations fournies par une table de correspondance (ligne e, Table 4).
- construction des fragments d'arbre (ligne f) et sémantiques (ligne g) qui permettront de créer une grammaire sous forme de classes (méta-grammaire) et de tester la grammaire

noyau,
— compilation de la métagrammaire.

Pour évaluer la grammaire résultant de l'extraction, nous créons à partir de chaque exemple Verbnet et des rôles qui lui sont associés, une sémantique plate reflétant la sémantique de l'exemple (e.g., (1b) pour (1a)). Nous donnons ensuite en entrée au générateur GenI (Gardent & Kow, 2007b), ces représentations sémantiques couplées avec la grammaire lexicalisée extraite par notre procédure d'extraction et nous comparons la phrase générée par GenI avec l'exemple VerbNet initial en utilisant la métrique BLEU (Papineni *et al.*, 2002) qui positionne l'évaluation de qualité entre 0 et 1. Un score de 1 indique que la phrase générée est identique à l'exemple VerbNet initial.

## 6.2 Table de correspondance

La table de correspondance nous permet de définir les clés support à la factorisation des arbres. Elle permet la génération et l'exploitation d'un fragment d'arbre associé à chaque composant décrit par VerbNet.

Les données de correspondance répondent à trois besoins :
① nom de clé à positionner pour construire le fragment d'arbre comprenant le segment,
② segment à traiter par le biais d'un rôle ou bien sous forme de co-ancre,
③ segment impliquant la présence d'un élément le préfixant ou le suffixant.

Le composant *item_a* (Figure 3 et Table 3) est spécifié par VerbNet comme étant un `NP`, identifié sémantiquement comme `Theme`, et dépourvu de restriction syntaxique. On y associe une clé (`n` pour notre propre table de correspondance). L'arbre initial exploitant ce fragment verra son identifiant XMG bâti avec cette information (`n0VPrepPN` pour l'arbre généré à partir de la structure de *A fire raged in the mountains.*). Cela permet la gestion de la factorisation des arbres, donc la réexploitation des fragments et des arbres initiaux.

Le traitement des besoins ① et ② seront explicités dans la partie suivante, qui traite le processus d'extraction.

Le besoin ③, très spécifique, est lié aux restrictions syntaxiques. Par exemple, un segment syntaxique *NP* associé à l'information sémantique *Value* avec la restriction sémantique *+how_extract* donnera naissance à un nœud nommé *H* coancré par *how*, de catégorie *WRB*[3].
Cela implique la recherche du mot *how* lors de la fragmentation de la phrase exemple, la génération d'une co-ancre *how* dans l'arbre initial créé à partir des informations : sa production au sein de la méta-grammaire contenante.
La table de correspondance fournit l'appui nécessaire aux traitements.

86% des types de segment syntaxico-sémantique traités donneront lieu à la recherche d'un groupe nominal, d'un adverbe, d'un verbe ou encore d'un adjectif et à la création d'un nœud TAG correspondant. Pour eux, la table de correspondance suffit.
Le traitement est spécifique pour 27 types de nœuds sur les 192 potentiellement gérables, ce qui signifie la production de traits sémantiques adaptés lors de la génération des fragments.

3. WRB : catégorie de *how* au sein du jeu d'étiquettes du Penn Treebank Project, qui est le jeu que nous avons utilisé dans le cadre de cette expérience pour catégoriser les segments et leurs composants.

Le programme implémenté dans le cadre de l'expérience vient spécifiquement contraindre certains fragments.

### 6.3 Exemple d'extraction

La Table 4 illustre le processus d'extraction. Les trois premières lignes montrent la segmentation de l'exemple *A fire raged in the mountains* à partir des informations sémantiques (a), syntaxiques (b et c) en mots (d), ce guidés par les informations de correspondance (e). Ces étapes permettent la récupération de la structure XMG initiale pour le verbe. La ligne f donne le nom des classes XMG correspondantes et la ligne g montre la sémantique plate qui leur est associée.

| | | | | |
|---|---|---|---|---|
| a. *Structure sémantique* | Theme | Event | | Location |
| b. *Structure syntaxique* | NP | VERB | PREP | NP |
| c. *Restrictions syntaxiques* | ∅ | ∅ | ∅ | ∅ |
| d. *Exemple* | A fire | raged | in | the mountains |
| e. *Correspondance* | N *rôle* | V | P *rôle* | N *rôle* |
| | ↓ | ↓ | ↓ | ↓ |
| f. *Classes XMG* | NObjectB-11 | n0VPrepPN | redprainbs | PrepPNObjectA-12 |
| g. *Sémantique plate* | i :a_fire(w) | r :rage(e1)<br>r :Theme(e1 w) | m :in(e1 j) | z :the_mountains(j) |

Table 4 – Extraction d'un arbre intégrant un modificateur

La Figure 5 montre les fragments XMG résultant de notre procédure d'extraction. Cette procédure spécifie en outre que le fragment *NObjectB-11* (i) importe le fragment *Fragment Verbal importé* et (ii) est combiné par conjonction avec le fragment *PrepPNObjectA-12*. La compilation XMG produit l'arbre TAG n0VPrepPN (cf. Figure 1) qui sera associé dans le lexique de génération avec le verbe *to rage* ainsi qu'avec tous les verbes de la classe *entity specific modes being*[4].

Nous aboutissons à la génération automatique des arbres TAG, comme celui de n0VPrepPN, créé pour l'exemple présenté Table 4 de la classe *entity specific modes being*.
La sémantique plate permet de tester avec un réalisateur de surface que le texte est correctement généré à partir des arbres. Elle est formée en parallèle de la grammaire lexicalisée des verbes. Cela nous amène presque à la fin de cet article, aux tests, et donc aux résultats.

## 7 Résultats

VerbNet nous fournit 1415 exemples pour couvrir ses verbes. Certains exemples (91 sur 1415, soit 6%) servent à plusieurs structures. '*The matter seems in dispute.*' sert à couvrir une structure de la classe *seem-109.xml* et une autre de la classe *become-109.1.xml*. Dans la première il illustre une structure thème/verbe/attribut sentencieux[5], dans la seconde, une

4. *entity specific modes being* : cette classe regroupe des verbes permettant de spécifier une façon spécifique d'être pour des entités (ex : *burn*, *fizz*, *flower*)...

5. Extrait de VerbNet, *seem-109.xml* :
`example "The matter seems in dispute." → syntax Theme V {in} Attribute<-sentential>`

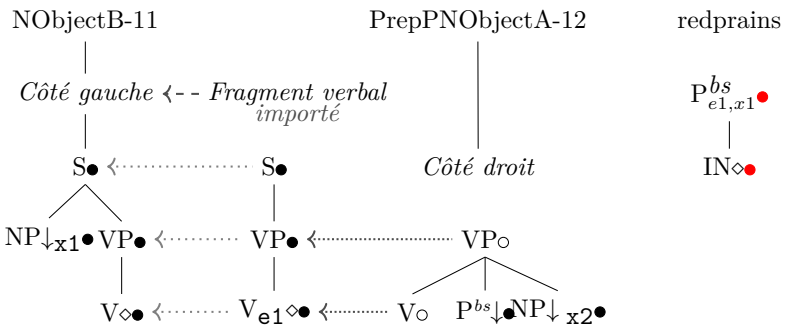



FIGURE 5 – composition de n0VPrepPN et fragment pour préposition

structure thème/verbe/résultat. Les exemples réutilisés peuvent l'être jusqu'à cinq fois.
De la même manière, inversement, des structures peuvent se voir illustrées par plusieurs exemples. Patient/verbe/with/co-patient est dans ce cas, elle a quatre exemples issus de quatre classes pour la démontrer :

— *The eggs mixed with the cream.*
— *My computer connected to his computer.*
— *Ice cream integrates with desserts.*
— *Company A associated with Company B.*

Nous avons 720 structures possibles, avec 514 qui sont illustrées par un seul et même exemple, les 206 autres qui sont dotées en moyenne de 4,78 exemples.
VerbNet est relativement riche, avec des phrase canoniques relativement courtes (5.93 mots en moyenne).

Sur notre corpus constitués des 1415 phrases-exemple de VerbNet, nous générons 1711 phrase à partir des représentations sémantiques plates dérivées des données VerbNet, de la grammaire extraite et du générateur GenI. Nous obtenons un score BLEU-4 [6] (utilisé dans 90% des cas, composés en moyenne de 6.3 mots) de 0.93, un BLEU-3 de 0.96 pour les 7% de trigrammes, et un BLEU-2 de 0.99 pour les 3% de bigrammes).

Si le traitement des phrases simples (ne contenant pas de subordonnées) est opérationnel, le traitement des subordonnées n'est pas pleinement fait. Les données de VerbNet ont permis d'y trouver des incohérences (par exemple, pour une structure déclarée NP V NP LEX$^{=up}$, l'exemple VerbNet fournissait un NP après le mot *up* syntaxiquement imposé comme clôturant la phrase).

Des corrections mineures ont donc été requises. Les caractéristiques des subordonnées sont fournies par VerbNet, elles ont donc été implémentées en tant que contraintes sur le groupe nominal décrit. Non seulement le fragment phrasique de la subordonnée n'a pas été créé pour que des éléments puissent s'y ancrer, mais, de fait, les incohérences n'ont pas été relevées. Concrètement, un arbre final existe attendant un fragment avec des caractéristiques propres

6. BLEU : Bilingual Evaluation Understudy : mesure de la proximité entre deux entrées, typiquement entre une sortie machine et une phrase en langue naturelle, BLEU-4 s'appuie sur des quadrigrammes (comparant les mots par succession de quatre), et n'est donc pas exploitable pour des phrases de trois mots où BLEU-3 peut être utilisé...

à la subordonnée attendue (ex : qu'elle doit être introduite par *that* et être à tel temps par rapport à la principale) ; on vérifie qu'elle peut se réaliser (si toute fois les fragments permettaient d'ancrer la partie adjointe) : par le biais du lexique, on fournit un groupe ayant les bonnes caractéristiques. Rien encore n'est prêt pour que ce groupe puisse être réalisé par le biais de la sémantique. Il y a quelques erreurs dans VerbNet. Quand un élément est fourni dans une structure comme élément lexical coancré et à la fois est indiqué par des informations syntaxiques que nous traitons comme des traits, cela nous donne une surgénération aberrante du type '*he asked what whether to go*' ou encore '*he asked her to to do*'.

La surgénération non souhaitable représente approximativement 7% des hypothèses produites, ce qui nous laisse un corpus correct de quasiment 1600 phrases pour nos 1415 références, et explique la hauteur du score BLEU.

Un faible pourcentage des surgénérations est lié au non usage de traits dans le cadre sémantique. Nous créons *ad hoc* le lexique, et '*he converted*' se voit généré aussi bien que '*he was converting*' pour une même sémantique alors que '*he was converting*' ne fait pas partie du corpus de référence. Les temps auraient du être fournis dans des traits au niveau sémantique, trouver leurs formes dans un lexique morphologique, et non figurer comme variation lexicale d'un même lemme verbal dans le lexique dédié.

Un certain nombre de phrases ne figurant pas dans le référentiel ont été générées tout en étant correctes pour la sémantique donnée. Par exemple, '*ellen told helen the situation*', '*ellen told the situation to helen*' ont été offertes en variation de '*ellen told helen about the situation*'.

Nous aurions du couvrir 100% du corpus. Il y a trois phrases que nous ne parvenons pas à regénérer telles qu'elles auraient du être (la diversité sémantique permet de générer une phrase autre, mais pas celle de départ). Comme elles sont peu nombreuses, nous pouvons nous y attarder.

- '*she says ' enchilada ' with a proper mexican accent*' → mauvaise lexicalisation, l'arbre est correct (perte d'une cote dans l'information générée).
- '*i broke the twig and the branch apart*' → l'arbre est correct, *the twig and the branch apart* forment ici un groupe nominal complet structurellement, notre procédure de séparation des groupes (nominaux/verbaux/prépositionnels) adaptée à VerbNet a échoué sur cet exemple, ce qui s'est traduit par une mauvaise lexicalisation puis une mauvaise regénération. [7].
- '*on his finger there sparkled a magnificent diamond*' → mauvaise gestion d'un sujet anonyme après un groupe prépositionnel (les sujets anonymes ont un traitement spécifique).

Nous profitons du deuxième exemple pour mettre à jour une faille expérimentale : les traits ont été sculptés sur mesure pour la phrase canonique type (sujet verbe complément(s)), le cas du sujet anonyme a été pris en charge sauf pour l'exemple pré-cité. Néanmoins, il serait nécessaire de pousser l'expérience à une véritable interprétation de ce qui est sujet : « All through the mountains raged a fire. » est un type de phrase représentant 0.8% du corpus. Sur ce 0.8% du corpus, l'accord sujet/verbe ne se ferait pas correctement : les traits ont été portés pour se synchroniser du premier groupe au second dans le cadre d'une structure sujet verbe suivis de compléments très largement dominante dans les données VerbNet.

7. lexicalisation : dans le cadre de cette expérience, pour les tests, les groupes nominaux ont été traité comme un tout, l'usage segmenté de leur contenu fait partie d'une autre expérience

Au final, à partir des structures canoniques de VerbNet, nous obtenons en génération 172 arbres initiaux composés de deux à trois fragments (côté gauche, côté droit, noyau verbal importé) couvrant, plus ou moins bien (le traitement des subordonnées étant à finaliser).

Pour ce qui est des fragments générés, les côtés gauches sont au nombre de 8, les côtés droits 123. Il faut ajouter aux côtés gauches le sujet anonyme, exploité par sept des 172 arbres initiaux au-delà de l'arbre initial réflexif évoqué Table 6. La classe XMG sujet anonyme a été fournie manuellement, comme les deux noyaux atomiques verbaux (nombre libre ou pluriel contraint).

Les fragments générés (voir Table 6) reposent sur l'exploitation de fragments produits manuellement (voir Table 5).

| **arbres initiaux** | **exemple** |
|---|---|
| noyau verbal | *There* ***happened*** *an accident* |
| sujet anonyme | ***There*** *happened an accident* |
| binaire : attribut d'objet | *thief* ***of*** *diamonds, the shelf* ***of*** *drinks* |

Table 5 – Arbres initiaux fournis

| **arbres initiaux finaux** | **nb** | **exemple** |
|---|---|---|
| réflexif | 1 | *It rains.* |
| unaire | 23 | *Susan was chitchatting.* ou *There happened an accident.* |
| binaire | 73 | *Carmen bought a dress.* |
| ternaire | 70 | *Ellen told Helen to come.* |
| quaternaire | 5 | *Carol cut the envelope open with the knife.* |
| **autres initiaux** | | |
| unaire | 11 | *Carol*, *nice*, *the*, *really* ... |
| unaire ou binaire | 2 | prépositions : *to*, *in*, *from*, adverbes :*down* |

Table 6 – Arbres initiaux générés

# 8 Conclusion

L'utilisation des ressources linguistiques permet, sans expertise linguistique propre à l'utilisateur, de générer un lexique verbal adossé à la grammaire TAG qui lui correspond. Plusieurs sources d'informations ont du être couplées pour valider les formes que peut prendre un verbe et pour définir finement les composants des groupes syntaxiques (pour information, il s'agissait de NLTK et Wordnet, que nous ne présentons pas ici car non centraux dans la méthode).
Nous pouvons attendre de notre système un taux catégoriel de phrases correctes, en l'état, à 92% (les 8% manquants étant les cas non générés, la surgénération et les soucis de traits).

Ce système est objectivement fonctionnel, exploitable.
L'usage de TAG permettant l'expansion de n'importe quel groupe, son formalisme permettant la manipulation, la continuation de la recherche est plus qu'envisageable.

Il serait nécessaire de gérer les traits sémantiques (eux aussi fournis dans verbnet) pour cadrer par exemple un *Agent*, le cas échéant, sur un trait type *animate* ou *organization* (ce qui

signifie de manière relativement claire « l'agent doit être animé ou être une organisation »). Couplés à la notion de rôles, ces informations parferaient le modèle et seraient utiles pour adosser à notre humble grammaire formes transitives et interrogatives. Sachant comment on pose une question dans une langue donnée, il est possible, en remplaçant un rôle par une préposition introductive, de donner une forme cohérente à une interrogation.

Au-delà des bases verbales (formes morphologiques et lexique liant verbes et arbres), de la grammaire produite, une architecture devrait être pensée pour faciliter l'emploi d'une production intermédiaire issue d'un tel processus car la production intermédiaire est elle-même très riche (informations détaillées touchant aux structures et rôles).

Pour en revenir à notre choix - profiter des larges ressources commençant à s'accumuler en TAL, nous avançons ce diagnostic qui peut sembler aller de soit après la lecture de ce bref article, mais que l'état de l'art ne montre pas être si courante. Le traitement de la langue est un domaine trop vaste pour pouvoir le révolutionner seuls, il est productif de maximiser l'usage de briques conçues par la communauté, dans une perspective visant parfois la "simple" meilleure compréhension fondamentale de la langue.

# Références


Cahill A. & Van Genabith J. (2006). Robust pcfg-based generation using automatically acquired lfg approximations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 1033–1040 : Association for Computational Linguistics.

Candito M.-H. (1996). A principle-based hierarchical representation of ltags. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, p. 194–199 : Association for Computational Linguistics.

Candito M.-H. (1998). Building parallel ltag for french and italian. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, p. 211–217 : Association for Computational Linguistics.

Carroll J. & Oepen S. (2005). High efficiency realization for a wide-coverage unification grammar. In *International Conference on Natural Language Processing*, p. 165–176 : Springer.

Chiang D. (2000). Statistical parsing with an automatically-extracted tree adjoining grammar. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, p. 456–463 : Association for Computational Linguistics.

Crabbé B., Duchier D., Gardent C., Le Roux J. & Parmentier Y. (2013). Xmg : extensible metagrammar. *Computational Linguistics*, **39**(3), 591–629.

Gardent C. (2006). Intégration d'une dimension sémantique dans les grammaires d'arbres adjoints. In *Actes de la conférence TALN 2006*, p. 149–158.

Gardent C. & Kallmeyer L. (2003). Semantic construction in feature-based TAG. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, p. 123–130 : Association for Computational Linguistics.

Gardent C. & Kow E. (2007a). GenI, un réalisateur basé sur une grammaire réversible. In *14e conférence pour le Traitement Automatique des Langues Naturelles-TALN 2007*, p. 10–p.

Gardent C. & Kow E. (2007b). A symbolic approach to near-deterministic surface realisation using tree adjoining grammar. In *45th Annual Meeting of the Association for Computational Linguistics-ACL 2007*, p. 328–335 : Association for Computational Linguistics.

Gyawali B. (2016). *Surface Realisation from Knowledge Bases*. PhD thesis, Universite de Lorraine.

Kipper K., Dang H. T., Palmer M. & others (2000). Class-based construction of a verb lexicon. *AAAI/IAAI*, **691**, 696.

Kipper K., Korhonen A., Ryant N. & Palmer M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, **42**(1), 21–40.

Montague R. (1973). The proper treatment of quantification in ordinary English. In *Approaches to natural language*, p. 221–242. Springer.

Narayan S. & Gardent C. (2012). Structure-driven lexicalist generation. In *24th International Conference in Computational Linguistics (COLING)*, p. 100–113.

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

Rogers J. & Vijay-Shanker K. (1994). Obtening Trees from Their Descriptions : An Application to Tree-Adjoining Grammars. *Computational intelligence*, **10**(4), 401–421.

Vijay-Shanker K. & Joshi A. K. (1988). Feature structures based tree adjoining grammars. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, p. 714–719 : Association for Computational Linguistics.

Wong Y. W. & Mooney R. J. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Annual Meeting-Association for computational Linguistics*, volume 1, p. 960.

Xia F. (2001). *Automatic grammar generation from two different perspectives*. phdthesis, University of pennsylvania.

# Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique

Yoann Dupont[1, 2]
(1) Laboratoire Lattice (CNRS, ENS, Université Sorbonne Nouvelle, PSL Research University, USPC)
1 rue Maurice Arnoux, 92120 Montrouge
(2) Expert System France, 207 rue de Bercy, 75012 Paris
yoa.dupont@gmail.com

RÉSUMÉ

Dans cet article, nous explorons divers traits proposés dans la littérature afin de fournir un détecteur d'entités nommées pour le Français appris automatiquement sur le French Treebank. Nous étudions l'intégration de connaissances en domaine, l'apport de la classification des verbes, la gestion des mots inconnus et l'intégration de traits non locaux. Nous comparons ensuite notre système aux récents réseaux de neurones.

ABSTRACT

**Feature exploration for French Named Entity Recognition with Machine Learning**

In this article, we explore features that have been described in the litterature to learn a named entity recognizer on the French TreeBank. We study the use of in-domain knowledge, the benefit of classifying verbs, the handling of unknown words and non-local features. We then compare our system to recent neural networks.

MOTS-CLÉS : Reconnaissance d'entités nommées, French Treebank, Apprentissage automatique, CRF, réseaux de neurones.

KEYWORDS: named entity recognition, French Treebank, machine learning, CRF, neural networks.

## 1 Introduction

La reconnaissance d'entités nommées (REN) est une tâche importante du TAL qui sert généralement de point de départ à d'autres tâches telles que l'extraction de relations (Bunescu & Mooney, 2005), l'entity linking, la résolution de coréférence (Hajishirzi *et al.*, 2013). De nombreux travaux ont été réalisés dans le domaine de l'apprentissage automatique sur l'anglais, mais assez peu sur d'autres langues, en particulier le français, pour lequel on peut citer Galliano *et al.* (2009); Gravier *et al.* (2012); Sagot *et al.* (2012). Les interactions entre les différents traits que l'on peut intégrer dans ces systèmes ont déjà fait l'objet d'études sur l'anglais Ratinov & Roth (2009); Tkachenko & Simanovsky (2012), mais à notre connaissance aucune n'a été faite sur le français. Ici, nous explorons divers traits utilisés dans la littérature afin d'en évaluer l'impact sur la REN pour le français. Nous partirons des traits proposés par Raymond & Fayolle (2010) sur le French Treebank (FTB) annoté en entités nommées (Sagot *et al.*, 2012), que nous enrichirons progressivement.

Nous étudions également comment organiser les ressources lexicales ainsi que la meilleure manière de les intégrer une fois cette classification faite, comme la classification des verbes et la gestion des mots inconnus.

## 2 La tâche et les corpus

Nous considérons la REN comme une tâche d'étiquetage de séquences. Les séquences à annoter sont les phrases d'un texte, où chaque élément est un *token*. Une entité pouvant correspondre à plusieurs tokens, nous avons utilisé le schéma d'annotation BIO (*Beginning*, *Inside Outside*), où un *token* est annoté `B-<Entité>` s'il est le premier *token* d'une entité, `I-<Entité>` s'il appartient à une entité sans être son premier élément et `O` s'il ne fait pas partie d'une entité. Nous avons principalement utilisé les CRF, particulièrement adaptés cette tâche, que nous comparons à des réseaux de neurones récents faisant partie des meilleurs systèmes sur l'anglais. Nous avons choisi le FTB annoté en entités nommées car il est le plus proche d'un point de vue typologique du corpus le plus utilisé pour l'anglais : le corpus CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003).

### 2.1 les CRF

Les *Conditional Random Fields* (CRF) (Lafferty *et al.*, 2001) sont des modèles graphiques probabilistes (Koller & Friedman, 2009; Gaussier & Yvon, 2011). Ils prennent en entrée un ensemble structuré d'éléments $x$ et fournissent en sortie un étiquetage structuré $y$ de cet ensemble. Ils sont dits discriminants car ils modélisent la probabilité conditionnelle d'un ensemble d'étiquettes $y$ selon une entrée $x$. Quand le graphe exprimant les dépendances entre étiquettes est linéaire, la distribution de probabilité d'une séquence d'annotations $y$ selon une séquence observable $x$ est donnée par :

$$p(y|x) = \frac{1}{Z(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right] \quad (1)$$

où $Z(x)$ est un facteur de normalisation dépendant de $x$ et où les $K$ traits $f_k$ sont des fonctions à valeur dans $\{0, 1\}$ fournies par l'utilisateur. Les poids $\lambda_k$ associés aux différents traits $f_k$ sont les paramètres du modèle déterminés par l'apprentissage. L'une des implémentations les plus efficaces des CRF linéaires est fournie par Wapiti [1] (Lavergne *et al.*, 2010), qui implémente la plupart des algorithmes d'entraînement couramment utilisés en plus de permettre la sélection de traits pertinents.

Dans nos expériences, nous nous concentrerons sur le French Treebank (FTB) (Abeillé *et al.*, 2003) annoté en entités nommées (Sagot *et al.*, 2012).

### 2.2 Le FTB

Le French Treebank, ou FTB (Abeillé *et al.*, 2003), est un recueil de phrases issues du journal Le Monde de 1989 à 1995 annotées en constituants. Sa taille est de 12351 phrases pour 350931 . Dans le cadre de cette expérience, nous avons utilisé sa version annotée en entités nommées fournie par

1. disponible depuis : https://github.com/Jekub/Wapiti

Sagot *et al.* (2012), dont les caractéristiques principales sont données dans le tableau 1. On y distingue sept types d'entités principaux : *Company* (les entreprises), *Location* (les lieux tels que les villes ou les pays), *Organization* (les organisations à but non lucratif), *Person* (les personnes réelles), *Product* (les produits), *FictionCharacter* (les personnages fictifs, de série TV ou bande dessinée par exemple) et finalement les *PointOfInterest* (les points d'intérêt tels que l'Opéra). Le découpage du corpus suit le protocole entraînement–développement–test défini par Crabbé & Candito (2008).

| | Entraînement | Développement | Test |
|---|---|---|---|
| Phrases | 9881 | 1235 | 1235 |
| Entités | 9235 | 1271 | 1173 |

TABLE 1 – Une vue d'ensemble du FTB annoté en entités nommées

Typiquement, les systèmes de type CRF sur des tâches de reconnaissance d'entités nommées intègrent de la connaissance en domaine à l'aide de lexiques. De plus, près de 40% des entités du corpus de test sont absentes du corpus d'entraînement. Nous commencerons donc par détailler l'intégration de ce type de connaissances dans notre système.

# 3 Intégrer de la connaissance en domaine

Nombre de systèmes par apprentissage intègrent des connaissances d'une façon ou d'une autre, ceux n'en intégrant aucune peinant à être compétitifs, comme l'a conclu Jungermann (2007). Cette connaissance peut autant être lexicale et morphologique (Raymond & Fayolle, 2010; Holat *et al.*, 2016), une représentation des mots calculée sur des volumes importants de données textuelles (Collobert & Weston, 2008; Ratinov & Roth, 2009; Lample *et al.*, 2016) ou le résultat de l'apprentissage joint de plusieurs tâches (Collobert & Weston, 2008; Luo *et al.*, 2015).

Une des premières sources de connaissances externes utilisée pour la REN est un ensemble de lexiques, généralement un par type de sortie. Les traits issus de l'utilisation d'un ensemble de lexiques, appelé « connaissances *a priori* », sont un point important de Raymond & Fayolle (2010), qui nous servira donc de point de départ. Ces traits sont en fait la combinaison de plusieurs éléments, que l'on peut rapprocher des motifs décrits par Holat *et al.* (2016). Ils sont générés en trois étapes :

1. Les connaissances *a priori* sont appliquées. Ici, chaque terme reconnu par un lexique est marqué avec l'identifiant de ce dernier.
2. Les mots dits « importants » (qui ont une forte information mutuelle avec une classe de sortie) sont laissés tels quels.
3. La partie du discours (*Part Of Speech*, POS) est utilisée pour les mots non reconnus dans les deux étapes précédentes.

Nous avons légèrement modifié ces traits ici : à la place des mots importants, nous avons créé des lexiques de termes déclencheurs pour chaque type d'entité principal. Il existe deux différences entre les mots importants et les termes déclencheurs. Là où les mots importants sont générés automatiquement, les lexiques de termes déclencheurs ont été constitués manuellement et enrichis à l'aide de noms communs récupérés dans le contexte proche des entités dans l'ensemble d'entraînement. La nature de l'information est également différente : pour un mot important, sa forme fléchie est utilisée, alors que nous utilisons l'identifiant du lexique est utilisé pour un terme déclencheur. L'utilisation

des formes fléchies pour les mots importants pose à notre avis deux problèmes. Premièrement cette liste est figée et toute modification impose de réapprendre le modèle, alors qu'un terme déclencheur peut simplement être ajouté à la liste correspondante. De plus, il n'est pas garanti que l'ensemble des mots importants présents dans l'ensemble d'apprentissage soit exhaustif. Si un mot important est absent du corpus d'apprentissage, les CRF seront incapables de l'utiliser pendant l'annotation. À l'inverse, les termes absents des lexiques peuvent être rajoutés au fur et à mesure, donnant aux CRF des informations qu'ils peuvent utiliser.

Le tableau 2 présente des traits générés par la procédure précédente. Les lignes commençant par « l>X » signifient que « X » a été utilisé pour les mots qu'aucun lexique n'a reconnu. Nous voulions évaluer l'utilisation de deux ressources en plus du POS, l'une plus précise — les mots — et l'autre plus générale — le chunking (Abney, 1991). L'intuition est que les mots permettent d'avoir des contextes plus forts, tandis que le chunking permet une plus grande généralisation — les entités nommées correspondant généralement à des chunks nominaux ou prépositionnels.

| **Mots** | la | société | Warner | fondée | par | les | frères | Warner |
|---|---|---|---|---|---|---|---|---|
| **lexiques** | ∅ | company.trigger | last-name | ∅ | ∅ | ∅ | ∅ | last-name |
| **l>mots** | la | company.trigger | last-name | fondée | par | les | frères | last-name |
| **l>POS** | DT | company.trigger | last-name | ADJ | PRP | DET | NC | last-name |
| **l>chunks** | NP | company.trigger | last-name | AP | B-PP | I-PP | I-PP | last-name |

TABLE 2 – Exemple de traits générés depuis un répertoire de lexiques.

Nous nous concentrerons ici sur la gestion des termes ambigus dans la source de connaissance, c'est-à-dire ceux qui apparaissent dans au moins deux lexiques différents. Nous évaluons pour cela deux méthodes de gestion de ces ambigüités, lesquelles ne figuraient pas dans les travaux originaux mais qui peuvent causer de grandes différences de résultats, comme nous le verrons dans la section 4.2. Il n'est pas rare qu'un terme puisse être ambigu, dans le sens où il apparaît dans plusieurs lexiques. C'est par exemple le cas de « Paris », qui peut référer à une ville (lieu) ou à un prénom (personne). Dans un tel cas, deux possibilités s'offrent à nous. La première consiste à effectuer une analyse ambigüe, où chaque terme reconnu par plusieurs lexiques se verra attribuer plusieurs classes. Ce type d'analyse a l'avantage de distinguer les termes sûrs de ceux ambigus et permet donc à l'algorithme d'effectuer une analyse plus fine, mais elle a également comme inconvénient que les ambigüités peuvent ne pas être observées à l'apprentissage, laissant le système démuni en phase d'annotation. Une seconde approche consiste à établir une relation d'ordre sur les lexiques, notée $>$, avec $x > y$ se décrivant comme « $x$ est plus prioritaire que $y$ ». Prenons deux lexiques $x$ et $y$ ainsi qu'un terme $t$ tels que $t \in x \cap y$ et $x > y$. Lorsque nous rencontrons le terme $t$ dans notre corpus, ce dernier prendra alors systématiquement la classe associée à $x$. Cette approche a l'avantage de ne laisser aucune ambigüité et de fournir des traits plus simples et moins silencieux au CRF, même si ces derniers sont moins précis. Par la suite, nous appellerons ces connaissances *a priori*, lorsqu'elles sont classées et triées, un *répertoire* de lexiques, dont nous étudierons l'intégration dans un CRF.

La clasification utilisée pour la REN sur le FTB est donnée dans la figure 1. Nous avons également utilisé la classification des verbes de Dubois & Dubois-Charlier (1997), qui définit en fait deux classifications des verbes différentes : une générique (communication, don/privation, auxiliaires, etc…) et une sémantique (humain, animé, non-animé, etc…). Nous n'avons pas obtenu de meilleurs résultats en intégrant cette classification dans le *répertoire*, raison pour laquelle elle n'y figure pas.

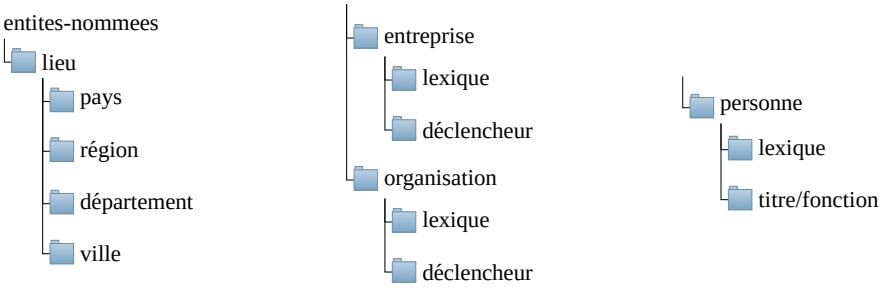



FIGURE 1 – clasification des lexiques utilisée dans les systèmes présentés. Plus un lexique est haut, plus ce dernier est prioritaire.

# 4 Expériences

## 4.1 Résultats préliminaires sur le FTB

Nous avons utilisé le répertoire décrit dans la section 3 pour établir les priorités de nos lexiques en cas d'ambigüité. Nous avons appliqué les lexiques en ignorant la capitalisation des mots, cette information étant plus pertinente à utiliser comme trait dans notre CRF. En effet, ne pas considérer la casse au moment d'appliquer les lexiques maximise leur couverture, ajouter les traits de capitalisation dans le CRF lui permettant alors d'apprendre à distinguer les reconnaissances bruitées.

L'ensemble des traits décrits dans ce papier ont été générés dans une fenêtre de $[-2, 2]$ autour du token courant. Notre CRF *étalon* utilise les mots, les lexiques comme traits spécifiques (un trait booléen par lexique), des préfixes et suffixes jusqu'à une taille de 5 caractères ainsi que divers traits booléens comme des informations de capitalisation. Dans notre expérience visant l'intégration du répertoire de lexiques, nous avons utilisé la configuration permettant d'obtenir les meilleurs résultats dans la section 4.2. Nous avons également intégré successivement les préfixes et suffixes des mots, les noms voisins ainsi que le verbe suivant (plus précisément, le premier verbe non suivi d'un autre verbe afin de ne pas capturer les auxiliaires des participes passés). Les classes de verbes sont celles définies par Dubois & Dubois-Charlier (1997). Nous avons utilisé la classe générique et la classe sémantique. Si un verbe avait plusieurs acceptions ayant des classes différentes, nous avons choisi la première, qui correspond à celle du verbe dans son sens premier.

Les résultats sont détaillés dans le tableau 3. Les expériences (a) à (c) montrent l'utilisation de mots et du répertoire uniquement. De façon assez prévisible, les mots donnent une très forte précision mais un mauvais rappel, tandis que les POS donnent un meilleur rappel. Les chunks semblent être à mi-chemin entre ces deux informations en termes de qualité, ce qui peut paraître étonnant étant donné qu'il s'agit d'une information plus générale que le POS. Cela vient du fait que la plupart des entités nommées correspondent à un chunk nominal sans le déterminant, la fin d'une entité correspondant à la fin d'un chunk nominal. La majorité des erreurs faites dans (c) sont sur des entités de taille 1 absentes des lexiques, où l'information du chunk n'est donc pas utilisable. Les expériences (d) et (e) montrent que le POS semble avoir un léger avantage par rapport au chunking en termes de qualité, particulièrement en termes de rappel. La combinaison des différentes expériences de (a) à (c) n'a pas donné d'amélioration significative du modèle, nous avons donc utilisé l'expérience (b) comme base

| Expérience | Précision | Rappel | F-mesure |
|---|---|---|---|
| CRF *étalon* | 85.89 | 76.88 | 81.13 |
| a. o/mots | **89.42** | 69.20 | 78.02 |
| b. o/POS | 85.4 | 76.88 | 80.92 |
| c. o/chunking | 88.95 | 74.83 | 81.28 |
| d. (b) +préfixes/suffixes | 86.48 | 78.58 | 82.34 |
| e. (c) +préfixes/suffixes | 87.26 | 77.73 | 82.22 |
| f. (d) +noms voisins | 85.86 | 78.75 | 82.15 |
| g. (d) +prochain verbe (forme) | 86.21 | **78.92** | 82.41 |
| h. (d) +classes prochain verbe | 85.89 | **78.92** | 82.26 |
| (f) + (g) | 86.03 | 78.84 | 82.28 |
| (f) + (h) | 86.77 | **78.92** | **82.66** |

TABLE 3 – Les premiers résultats obtenus sur le FTB. En gras sont marqués les meilleurs scores pour la colonne.

pour les autres expériences. L'ajout des verbes et classes de verbes n'a pas donné d'amélioration significative et donnait même plus souvent lieu à une dégradation des résultats.

De nombreux termes présent dans le répertoire sont ambigus, dans le sens où ils peuvent apparaître dans plusieurs lexiques. Dans la section suivante, nous détaillerons comment nous avons géré ces ambiguïtés et comment cette gestion peut influer sur la qualité finale du modèle.

## 4.2 Gestion de l'ambiguïté des lexiques

Comme dit dans les parties précédentes, le choix des priorités quant aux différents types du répertoire est capital. Dans cette section, nous détaillerons les différences de résultats que ces changements peuvent amener. À cet effet, nous avons simplement évalué les différents ordonnancements ainsi qu'en effectuant une analyse ambigüe, où plusieurs éléments d'un répertoire peuvent reconnaître un même mot ou groupe de mots. Ici, nous évaluons l'impact sur les résultats obtenus sur le FTB en changeant l'ordre de priorité des lexiques au moment de générer les traits relatifs aux répertoires. Nous n'avons pas inclus l'influence entre organisation et entreprise car ces dernières n'avaient aucun terme en commun. L'analyse dite ambigüe consiste à expliciter l'ensemble des ambigüités présentes dans les lexiques, ce qui se fait ici en effectuant la concaténation des différents lexiques ayant reconnu un mot. Ainsi, pour chaque mot du texte, nous pouvons récupérer l'ensemble des classes auxquelles il peut appartenir et ainsi avoir des termes reconnus de façon non-ambigüe ainsi que des termes reconnus de façon ambigüe. Le contexte doit alors être utilisé pour trouver la classe la plus appropriée. Le tableau 4 détaille les résultats obtenus en modifiant l'ordre de priorité des lexiques ainsi qu'en effectuant une analyse ambigüe. Comme nous pouvons le remarquer, cet ordre influe de façon significative sur la qualité globale du résultat. Nous pouvons cependant déduire certaines tendances quant à l'ordre des lexiques. Par exemple, le lexique des lieux doit être prioritaire sur celui des personnes, et il en va de même pour le lexique des organisations et entreprises. Notre lexique des personnes ayant un bruit assez important, lui donner une priorité faible permet de privilégier les autres lexiques, plus sûrs. L'influence est moindre entre les lieux et les organisations/entreprises, car il y a peu d'intersection entre ces derniers.

| Expérience | Précision | Rappel | F-mesure |
|---|---|---|---|
| $P > C\&O > L$ | 85.98 | 76.79 | 81.13 |
| $P > L > C\&O$ | 85.58 | 76.96 | 81.04 |
| $L > P > C\&O$ | 85.47 | 77.89 | 81.50 |
| $L > C\&O > P$ | 85.80 | 78.49 | 81.98 |
| $C\&O > P > L$ | 85.66 | 76.36 | 80.74 |
| $C\&O > L > P$ | **86.77** | **78.92** | **82.66** |
| Ambigüe | 85.39 | 76.79 | 80.86 |

TABLE 4 – Les résultats selon la priorité accordée aux différents lexiques. P : Person, L : Location, C&O : Company&Organization. En gras sont marqués les meilleurs scores pour la colonne.

L'analyse ambigüe est celle dont les performances sont les plus mauvaises, autant en termes de précision que de rappel. Les problèmes principaux de l'analyse ambigüe, en comparaison avec le meilleur système, sont d'abord le silence sur les lieux, suivi d'erreurs de type puis de frontières, les entités proposées tendant à être plus courtes. Si l'on observe les poids présents dans le CRF, ces derniers diffèrent peu entre traits ambigus et non-ambigus dans les cas les plus simples. Les traits ambigus quant à eux tendent à suivre l'ordre de priorité ayant donné le meilleur résultat, favorisant les *Company* et *Organization*, puis les *Location* et finalement les *Person*. Les silences sont généralement dus à des ambigüités non-observées dans le corpus d'apprentissage. Partant de ce constat, nous pouvons proposer une méthode afin d'estimer un ordre proche de l'optimal. Pour ce faire, nous apprenons un modèle où une analyse ambigüe a été effectuée. En recherchant dans ce modèle les poids attribués par le CRF pour résoudre les cas spécifiquement ambigus, il est possible d'avoir une idée des lexiques plus ou moins prioritaires. L'inconvénient d'une analyse ambigüe demeure dans la combinatoire des possibles ambigüités, qui fait que toutes ne peuvent pas toujours être observées. Les cas ambigus absents du corpus d'apprentissage ne pourront pas se voir attribuer un poids par le CRF, qui sera donc incapable d'en tirer profit à l'annotation.

De manière générale, l'inconnu est l'une des plus grandes sources d'erreurs pour les systèmes par apprentissage. La source principale d'inconnu demeure encore les formes fléchies non observées dans le corpus d'apprentissage, en particulier si ces dernières ne font partie d'aucun lexique. Dans la section suivante, nous souhaitons évaluer l'intégration des mots inconnus dans notre CRF.

## 4.3 Gestion des mots inconnus

Typiquement, lorsqu'un algorithme d'apprentissage automatique rencontre un mot inconnu, il recourt à son contexte et/ou à sa morphologie afin de trouver une classe pertinente. Le caractère inconnu d'un mot offre une information intéressante pour l'analyse de ce dernier : en effet, de nombreuses entités nommées peuvent être déclenchées dans le contexte d'un mot inconnu. Afin de donner aux CRF l'information du caractère inconnu d'un mot, nous avons extrait le lexique des mots du corpus d'apprentissage et supprimé ceux trop peu fréquents, laissant ainsi le lexique des mots considérés comme connus. Un mot inconnu est alors un mot absent de ce lexique. Une fois ce lexique constitué, il est possible d'ajouter de plusieurs façons l'information « mot inconnu ». La première consiste à intégrer ce lexique directement dans le répertoire en lui accordant la plus faible priorité : ainsi, les CRF pourront obtenir une vision contextualisée d'un mot inconnu et pourra ainsi

mieux le désambigüiser. Une seconde serait de rajouter un trait booléen « mot inconnu » afin de l'utiliser comme tout autre trait booléen dans le CRF. Cela permet également, pour les traits discutés dans la section 3, de rajouter une information supplémentaire pouvant s'ajouter aux informations déjà présentes : par exemple, un « nom propre inconnu » sera une information plus pertinente que simplement un « mot inconnu ». Dans nos expériences, seul l'ajout d'un nouveau trait, où les mots appartenant au lexique des mots inconnus étaient remplacés par "_unknown_", a permis d'obtenir des gains intéressants, les meilleurs résultats étant obtenus en considérant inconnus les mots apparaissant un maximum de 4 fois dans le corpus, après avoir testé l'ensemble des valeurs entre 1 et 5. Nous avons appelé ce trait "hapax4" (inspiré de Pécheux *et al.* (2015)), dans le tableau 5.

| Expérience | Précision | Rappel | F-mesure |
|---|---|---|---|
| a. CRF (section 4.2) | 86.77 | 78.92 | 82.66 |
| b. (a) +concat($mot_i$,$mot_{i+1}$) avec i ∈ {-2,1} | 87.59 | 79.52 | 83.36 |
| c. (a) +concat($hapax4_i$,$hapax4_{i+1}$) avec i ∈ {-2,1} | 88.15 | 79.95 | 83.85 |
| d. (c) +concat($hapax4_i$,o/$POS_0$) avec i ∈ {-2,-1,1,2} | **88.41** | **80.03** | **84.05** |
| (c) + mots inconnus = classe | 86.80 | 79.69 | 83.10 |

TABLE 5 – Les résultats en intégrant les mots inconnus. En gras sont marqués les meilleurs scores pour la colonne.

L'un des inconvénients principaux des CRF réside dans leur nature fondamentalement locale. Afin de pallier ce manque inhérent au modèle, nous avons testé deux approches pour améliorer la consistance des annotations au niveau global.

## 4.4 Consistance des annotations

L'un des défauts des systèmes par apprentissage vient du fait que l'inférence se fait de façon purement locale, une hypothèse d'indépendance étant faite afin de rendre le modèle calculable en pratique. Divers travaux ont été effectués afin de modéliser des dépendances non-locales et assurer la consistance des annotations, modélisant généralement des dépendances à l'échelle du document et du corpus (Krishnan & Manning, 2006; Ratinov & Roth, 2009). La première approche que nous utilisons est une simple propagation des annotations : après l'annotation du CRF, nous récupérons le lexique de chaque type d'entité retrouvé par le CRF. Pour les entrées figurant dans plusieurs lexiques, nous prenons alors l'entité qui lui est la plus fréquemment attribuée. Deux heuristiques peuvent alors être employées. La première consiste à n'appliquer ces lexiques que sur des portions non-annotées du texte, celles proposées par le CRF étant considérées comme meilleures de façon systématique. La seconde consiste à ne mettre à jour les annotations du CRF que dans le cas où une chaîne plus longue a été trouvée, aucun retypage des séquences de même taille n'étant effectué. Cette approche a l'avantage d'être très simple à mettre en place et d'être intuitivement sous-optimale : elle permet donc d'établir une référence des gains obtenables par cette approche. Nous avons également testé l'approche en deux passes à l'aide des traits *token majority* et *entity majority* tels que décrits dans Krishnan & Manning (2006); Mao *et al.* (2007); Ratinov & Roth (2009). Le trait *token majority* considère la classe majoritairement associée à chaque token indépendamment, en ignorant les le schéma BIO. Par exemple, si « Paris » apparait deux fois en tant qu'organisation et une en tant que lieu, alors le trait *token majority* attribuera la valeur « organisation » à toutes les occurrences de « Paris ». Le trait *entity majority* est analogue à *token majority*, mais considère les entités retrouvées

par le premier CRF. Si par exemple « Calvin Klein » a été annoté trois fois en tant qu'entreprise et deux fois en tant que personne, alors toutes les occurences de « Calvin Klein » seront annotées en tant qu'entreprise. Les égalités ont été résolues en utilisant les priorités établies dans la section 4.2 et les chevauchement ont été gérés selon la règle de « la première chaine la plus longue ».

| Expérience | Précision | Rappel | F-mesure |
|---|---|---|---|
| CRF (section 4.3) | **88.41** | 80.03 | 84.05 |
| heuristique1 | 87.89 | **82.34** | **85.02** |
| heuristique2 | 87.80 | 82.26 | 84.94 |
| deux passes | 87.72 | 81.66 | 84.58 |

TABLE 6 – Les résultats selon les différentes méthodes de propagation. En gras sont marqués les meilleurs scores pour la colonne.

Le tableau 6 résume les résultats obtenus avec les différentes méthodes de propagation. L'heuristique sans mise à jour des annotations du CRF donne des résultats sensiblement meilleurs que celle pouvant modifier les annotations du CRF. En observant les annotations, nous avons remarqué que cette différence était principalement due à des inconsistances d'annotation dans le gold standard pour certaines organisations, les résultats sur les autres entités étant meilleurs. Nous observons cependant une baisse de précision à l'échelle globale, cela vient des erreurs de bruits et des types ambigus (ex : lieux contre organisations) qui se propagent par cette méthode. L'approche en deux passes telle que décrite dans Krishnan & Manning (2006); Mao *et al.* (2007); Ratinov & Roth (2009) ne nous a pas offert d'amélioration supplémentaire par rapport à notre post-traitement plus simple.

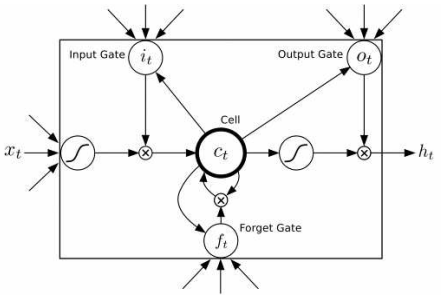
Depuis quelques années en particulier, les réseaux de neurones sont des concurrents sérieux aux CRF. Dans la section suivante, nous comparerons le système construit jusqu'ici avec l'une de ses variantes les plus efficaces, appelée le LSTM-CRF bidirectionnel.

## 5 Comparaison avec Bi-LSTM-CRF

Récemment, le domaine du TAL a vu un essor des réseaux de neurones *récurrents*. Ces derniers ont été créés pour traiter des séquences de données selon un principe simple : pour chaque élément d'une séquence, son résultat est injecté dans l'élément suivant de la séquence, permettant ainsi, de proche en proche, d'incorporer le contexte à un instant donné. Le réseau récurrent le plus simple est le réseau d'Elman (Elman, 1990), où la couche cachée d'un réseau communique avec elle-même. Cependant, ces réseaux n'arrivaient pas à modéliser des dépendances longue distance et leur apprentissage était très coûteux (Bengio *et al.*, 1994).

La couche cachée *Long Short-Term Memory* (Hochreiter & Schmidhuber, 1997), appelée par la suite LSTM, s'est distinguée par sa capacité à capturer des dépendances de longue portée. Il s'agit d'une couche cachée particulière d'un réseau d'Elman utilisant une mémoire interne capable de retenir les informations pertinentes, en ajoutant et oubliant des informations au fur et à mesure selon un système de portes permettant de n'activer un ajout ou un oubli qu'au moment opportun. Une illustration de la variante *peephole* (Gers & Schmidhuber, 2000) d'une cellule de LSTM ainsi que les formules des différentes portes sont données dans la figure 2. Nous comparerons notre CRF avec une variante de ce réseau faisant partie des systèmes les plus performants, appelée LSTM-CRF bidirectionnel (Bi-

LSTM-CRF) (Huang *et al.*, 2015; Lample *et al.*, 2016), qui sont parmi les systèmes état-de-l'art. Il s'agit d'une variante de réseau de neurones à laquelle est ajoutée un CRF, une de ses grandes forces étant sa capacité de généralisation (Augenstein *et al.*, 2017), leur permettant d'obtenir de très bons scores sur les entités inconnues. Dans cette partie, nous souhaitons effectuer une comparaison entre les différents CRF proposés ici et le réseau Bi-LSTM-CRF.




$$
\begin{aligned}
f_t &= \sigma(W_f \times x_t + U_f \times c_{t-1} + b_f) \\
i_t &= \sigma(W_i \times x_t + U_i \times c_{t-1} + b_i) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c \times x_t + b_c) \\
o_t &= \sigma(W_o \times x_t + U_o \times c_{t-1} + b_o) \\
h_t &= o_t \odot \sigma(c_t)
\end{aligned}
\tag{2}
$$

FIGURE 2 – représentation d'une cellule de LSTM et les formules des différentes portes. $\sigma$ représente la fonction sigmoïde. $\odot$ est le produit d'Hadamard. $\times$ est le produit matriciel. Illustration tirée de Graves *et al.* (2013)

Pour notre comparaison, nous avons utilisé l'implémentation de LSTM de Lample *et al.* (2016), qui a deux particularités. La première est qu'il utilise la variante de LSTM donnée sur la figure 2 et fusionne les porte d'oubli ($f_t$) et d'entrée ($i_t$). La seconde combine en réalité deux LSTM bidirectionnels. Un premier Bi-LSTM est employé pour donner une représentation des mots étant donné leur contexte gauche et droit dans la phrase, cela permet donc de capturer des informations de nature plus syntaxique. À cette représentation contextuelle des mots est alors concaténée une seconde, endogène, à l'aide d'un Bi-LSTM au niveau des caractères d'un mot, permettant de modéliser des informations relatives à la morphologie dudit mot. Ainsi, pour chaque mot, ce réseau modélise des informations à la fois syntaxiques et morphologiques, le rendant très intéressant pour des tâches telles que la REN. Un schéma de ces deux niveaux de représentation est donné dans la figure 3.

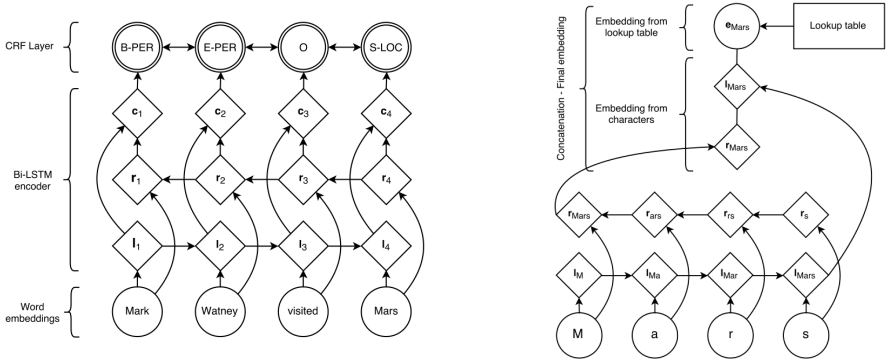



FIGURE 3 – À gauche : le Bi-LSTM-CRF pour une phrase. À droite : Bi-LSTM pour un mot. Illustrations tirées de Lample *et al.* (2016)

## 5.1 Expériences de comparaison

Pour cette comparaison, nous avons utilisé le Bi-LSTM-CRF proposé par Lample *et al.* (2016) [2]. Afin d'obtenir une comparaison complète, nous l'avons tout d'abord entraîné sans préentraînement puis en utilisant des représentations préappris sur des corpus issus du web [3]. Ils ont été appris à l'aide de word2vec (Mikolov *et al.*, 2014) sur le corpus FrWac du projet WaCky (Baroni *et al.*, 2009) et sur un dump de Wikipedia français. Ces derniers donnant des résultats significativement moins bons que ceux appris sur FrWac, nous ne les détaillerons pas ici. Les représentations appris sur le corpus FrWac ont une taille de 200, les mots apparaissant moins de 100 fois ont été considérés comme inconnus.

Pour notre Bi-LSTM-CRF, nous avons utilisé des vecteurs de taille 200 pour les mots et de taille 25 pour les caractères. Nous avons utilisé un dropout de 0,5. Les modèles ont été entraînés pendant 50 itérations selon une descente de gradient stochastique, le modèle final étant celui ayant maximisé la F1-mesure sur les entités dans le corpus de développement. Nous avons également évalué le gain obtenu en utilisant les traits de Raymond & Fayolle (2010) en tant qu'information supplémentaire de notre réseau, en leur attribuant une taille de 32, taille donnant les meilleurs résultats dans nos expériences. Les résultats comparatifs entre les CRF et les LSTM sont donnés dans le tableau 7.

| Système | Connues | | | Inconnues | | | Global | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| CRF *étalon* (section 4.1) | 95.04 | 92.34 | 93.67 | 68.68 | 53.53 | 60.17 | 85.89 | 76.88 | 81.13 |
| CRF (section 4.3) | **97.21** | 93.90 | 95.53 | 72.63 | 59.20 | 65.17 | **88.41** | 80.03 | 84.05 |
| +heuristique1 (section 4.4) | 96.83 | **95.46** | **96.14** | 72.46 | 62.53 | 67.13 | 87.89 | 82.34 | 85.02 |
| LSTM-CRF (*base*) | 96.10 | 94.33 | 95.20 | 64.21 | 54.18 | 58.77 | 84.53 | 78.33 | 81.31 |
| + o/POS | 95.95 | 94.04 | 94.99 | 70.13 | 59.13 | 64.27 | 86.56 | 80.20 | 83.26 |
| LSTM-CRF (FrWac) | 96.25 | 94.61 | 95.42 | 69.44 | 60.81 | 64.84 | 86.30 | 81.14 | 83.64 |
| + o/POS | 96.11 | 94.61 | 95.35 | **74.50** | 64.45 | 69.12 | 88.16 | 82.59 | 85.29 |
| + heuristique1 | 95.98 | 94.89 | 95.44 | 73.09 | **67.45** | **70.16** | 87.23 | **83.96** | **85.57** |
| Dupont & Tellier (2014) | ? | ? | ? | ? | ? | ? | 86.38 | 80.30 | 83.23 |

TABLE 7 – comparaison entre les différents CRF et LSTM-CRF. En gras sont marqués les meilleurs scores pour la colonne.

Les seuls résultats à notre connaissance sur la reconnaissance d'entités nommées sur le FTB sont ceux de Dupont & Tellier (2014), qui utilisent un CRF dont les traits sont proches de ceux de notre CRF *étalon*. Malgré la qualité inférieure de notre CRF *étalon* par rapport au leur, l'utilisation des traits détaillés ici nous a permis d'obtenir une qualité finale supérieure, ce qui montre bien leur pertinence. Nos Bi-LSTM-CRF intégrant des informations extérieures, autant sous la forme des traits détaillés dans la section 3 que de représentations préappris, témoignent également de l'efficacité des réseaux de neurones récurrents.

Les deux systèmes ont des f-mesures globales comparables, avec cependant un léger avantage pour le Bi-LSTM-CRF, qui se distingue surtout sur les entités inconnues où il obtient une qualité significativement supérieure à celle du CRF, résultats cohérents avec ceux décrits par Augenstein *et al.* (2017). L'ajout du répertoire de lexiques a permis au réseau de neurones d'obtenir une représentation plus

2. disponible à l'adresse : https://github.com/glample/tagger
3. disponibles à l'adresse : http://fauconnier.github.io

générale, en témoigne le gain de 5 à 6 points de f-mesure sur les entités inconnues, et de moins de 0,5 points sur les entités connues. L'ajout de règles de post-traitement simples a permis l'amélioration des deux meilleurs modèles respectifs, le CRF ayant plus bénéficié de ce gain que le LSTM-CRF. Cette différence s'explique par le côté plus précis du CRF, qui aura plus tendance à n'annoter une entité que dans un contexte sûr, une même entité n'étant alors annotée qu'à certains endroits. Une autre source de gain du Bi-LSTM-CRF vient des représentations préentraînés sur le corpus FrWac. Il est possible d'ajouter des représentations des mots dans un CRF, typiquement via l'entraînement de clusters de Brown (Brown *et al.*, 1992). Nous avons intégré 1000 clusters de Brown appris sur un dump de Wikipédia français, mais ces derniers n'ont pas amélioré nos résultats. Nous n'avons pas pu les entraîner sur le corpus FrWac, le coût en temps étant prohibitif.

L'ajout de la consistance des annotations en tant que post-traitement simple n'améliore pas significativement les résultats obtenus par les Bi-LSTM-CRF, en raison d'une forte baisse de la précision (-0.93). Cela vient du fait que le réseau de neurones est un système ayant tendance à être plus bruyant qu'un CRF, qui lui aura plus tendance à être silencieux. Cela se remarque également dans l'amélioration du rappel, moins importante pour le Bi-LSTM-CRF (+1.37) que pour le CRF (+2.31), autant sur les entités connues qu'inconnues. Cela suggère que le CRF a besoin d'un contexte plus fort pour annoter une entité et que Bi-LSTM-CRF a tendance à être plus consistant dans ses annotations.

# 6 Conclusion

Dans cet article, nous avons exploré diverses méthodes pour intégrer de la connaissance en domaine ainsi que des règles basiques dans un CRF. Nous avons utilisé comme point de départ les traits définis par (Raymond & Fayolle, 2010), que nous avons ensuite progressivement enrichis avec de nouvelles informations. Nous avons comparé cette méthode avec celle, classique, de l'ajout de lexiques à l'aide de traits booléens dissociés et avons constaté que leur combinaison améliorait les résultats car elle permet de contextualiser l'information des lexiques. Nous avons également étudié la classification des verbes ainsi que la meilleure manière des les intégrer au CRF. Nous avons aussi exploré la gestion des mots inconnus, leur ajout a permis d'obtenir une améliotation dans certaines conditions. Nous avons ensuite comparé notre CRF avec un Bi-LSTM-CRF ayant obtenu de meilleurs résultats avec moins de traits. L'une des premières tâches à effectuer serait d'ajouter les traits du CRF manquant au Bi-LSTM-CRF.

Nous prévoyons d'enrichir les lexiques que nous avons constitués ainsi que les traits de notre CRF. Pour cela, nous prévoyons d'ajouter des règles à notre CRF, autant en tant que traits qui seront pondérés à l'apprentissage qu'en tant que contraintes afin de forcer, ou d'interdire, le déclenchement d'une entité ou d'une annotation incohérente. Une de ces contraintes serait d'appliquer un poids négatif arbitrairement grand aux transitions inconsistantes, comme deux étiquettes `I-<Entité>` successives de deux entités différentes. Une autre idée serait d'utiliser des systèmes comme mXs (Nouvel *et al.*, 2013) qui disposent de nombreux lexiques et règles d'annotations que nous pourrions intégrer directement en tant que traits dans notre CRF. Nous comptons également explorer d'autres types de réseaux de neurones comme les *Gated Recurrent Units* (Cho *et al.*, 2014), plus légers, ainsi que de remplacer le Bi-LSTM au niveaux des caractères par une convolution.

Nous voulons également vérifier notre approche en l'appliquant sur d'autres langues et corpus comme le CoNLL (Tjong Kim Sang & De Meulder, 2003), où les systèmes actuels sont des plus compétitifs (Passos *et al.*, 2014; Luo *et al.*, 2015; Lample *et al.*, 2016).

# Références


ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.

ABNEY S. (1991). Parsing by chunks. In *Principle-Based Parsing*, p. 257–278 : Kluwer Academic Publishers.

AUGENSTEIN I., DERCZYNSKI L. & BONTCHEVA K. (2017). Generalisation in named entity recognition : A quantitative analysis. *arXiv preprint arXiv :1701.02877*.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3), 209–226.

BENGIO Y., SIMARD P. & FRASCONI P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, **5**(2), 157–166.

BROWN P. F., DESOUZA P. V., MERCER R. L., PIETRA V. J. D. & LAI J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, **18**(4), 467–479.

BUNESCU R. C. & MOONEY R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference HLT/EMNLP*, p. 724–731 : ACL.

CHO K., VAN MERRIËNBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, p. 1724–1734.

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of ICML*, p. 160–167 : ACM.

CRABBÉ B. & CANDITO M. H. (2008). Expériences d'analyse syntaxique statistique du français. In *Actes de TALN'08*.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). Synonymie syntaxique et classification des verbes français. *Langages*, p. 51–71.

DUPONT Y. & TELLIER I. (2014). Un reconnaisseur d'entités nommées du français. In *TALN 2014*, p. 40–41.

ELMAN J. L. (1990). Finding structure in time. *Cognitive science*, **14**(2), 179–211.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, volume 9, p. 2583–2586.

GAUSSIER É. & YVON F. (2011). *Modèles statistiques pour l'accès à l'information textuelle*. Lavoisier.

GERS F. A. & SCHMIDHUBER J. (2000). Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS IJCNN'00*, volume 3, p. 189–194 : IEEE.

GRAVES A., MOHAMED A.-R. & HINTON G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP 2013*, p. 6645–6649 : IEEE.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the french language. In *LREC-Eighth international conference on Language Resources and Evaluation*, p. n/a.

HAJISHIRZI H., ZILLES L., WELD D. S. & ZETTLEMOYER L. S. (2013). Joint coreference resolution and named-entity linking with multi-pass sieves. In *EMNLP*, p. 289–299.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

HOLAT P., TOMEH N., CHARNOIS T., BATTISTELLI D., JAULENT M.-C. & MÉTIVIER J.-P. (2016). Fouille de motifs et CRF pour la reconnaissance de symptômes dans les textes biomédicaux. In *Proceedings of TALN*.

HUANG Z., XU W. & YU K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv :1508.01991*.

JUNGERMANN F. (2007). Named entity recognition without domain-knowledge using conditional random fields. In *Workshop Notes of the Machine Learning for Natural Language Processing Workshop*, p. 16–17 : Citeseer.

KOLLER D. & FRIEDMAN N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press.

KRISHNAN V. & MANNING C. D. (2006). An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st COLING and the 44th annual meeting of the ACL*, p. 1121–1128 : ACL.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, p. 282–289.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings of ACL'2010*, p. 504–513 : Association for Computational Linguistics.

LUO G., HUANG X., LIN C.-Y. & NIE Z. (2015). Joint named entity recognition and disambiguation. In *Proc. EMNLP*.

MAO X., XU W., DONG Y., HE S. & WANG H. (2007). Using non-local features to improve named entity recognition recall. In *PACLIC*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2014). word2vec.

NOUVEL D., ANTOINE J.-Y., FREEBURGER N. & SOULET A. (2013). Fouille de règles d'annotation partielles pour la reconnaissance des entités nommées. In *Proceedings of TALN*.

PASSOS A., KUMAR V. & MCCALLUM A. (2014). Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv :1404.5367*.

PÉCHEUX N., ALLAUZEN A., LAVERGNE T., WISNIEWSKI G. & YVON F. (2015). Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles CRF. In *Conférence sur le Traitement Automatique des Langues Naturelles*.

RATINOV L. & ROTH D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th CoNLL*, p. 147–155 : ACL.

RAYMOND C. & FAYOLLE J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *TALN'10*.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, p. 142–147 : ACL.

TKACHENKO M. & SIMANOVSKY A. (2012). Named entity recognition : Exploring features. In *KONVENS*, p. 118–127.

# Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires

Claire Wolfarth[1]
(1) Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France
Claire.Wolfarth@univ-grenoble-alpes.fr

RESUME

L'émergence des corpus scolaires et la volonté d'outiller ces corpus spécifiques font apparaitre de nouvelles problématiques de recherche pour le traitement automatique des langues (TAL). Nous exposons ici une recherche qui vise le traitement de productions d'apprenants en début d'apprentissage de l'écriture, en vue d'une annotation et d'une exploitation ultérieure. À cette fin, nous proposons d'envisager cette étape comme une tâche d'alignement entre la production de l'apprenant et une normalisation produite manuellement. Ce procédé permet d'augmenter significativement les scores d'identification des formes et lemmes produits et améliore les perspectives d'annotation.

ABSTRACT

**To align production and normalization : first approach to study school learner's writings**
The emergence of school corpora and the will to provide tools for such specific corpora bring to light new research issues for natural language processing (NLP). We will expose a research which aims at correcting early learners' written production with the purpose of annotating and exploiting it at a later stage. We are putting forward to consider this stage as an alignment task between the learners' written and a manually produced 'normalized version' of it. This method increases significantly lemmas and forms identification task results and improves annotation possibilities.

MOTS-CLES : corpus scolaires, alignement, normalisation.
KEYWORDS: primary school corpus, alignment, normalization.

## 1 Production et normalisation : contexte de notre approche

Si l'étude de l'apprentissage de la lecture a permis l'outillage de son enseignement (Lector & Lectrix[1] ; Lectorino & Lectorinette[2] ; Bianco, 2017[3]), de tels outils n'existent pas encore pour l'apprentissage de l'écriture. À cette fin, plusieurs corpus scolaires de productions d'apprenants en langue française ont été constitués. Citons le corpus élaboré par M.-L. Elalouf (2005), rassemblant quelques 500 textes dans huit classes de CM2 et de 6e ; le corpus élaboré par H. Andersen, C.

[1] Cèbe, S., & Goigoux, R. (2009). Lector & [et] Lectrix: 20 posters pour l'étude des textes en collectif: CM1, CM2, 6e, Segpa. Retz.
[2] Goigoux, R., & Cèbe, S. (2013). Lectorino & Lectorinette CE1-CE2: apprendre à comprendre des textes narratifs. Retz.
[3] Bianco, M. (2017). Comment enseigner la compréhension ? *Enseigner à l'école primaire*. Hatier.

Leblay et E. Auriac-Slusarcyk (2010), contenant plusieurs centaines de récits et comptes-rendus scientifiques, recueillis en CE2, CM2, 6e et 4e ; le corpus recueilli par C. Garcia-Debanc et K. Bonnemaison (2014), de près de 400 textes autour d'une tâche de cohésion textuelle ; le corpus élaboré par J. David et C. Doquet (2016), rassemblant plus de 800 productions d'apprenants à l'école primaire et le corpus sur lequel nous travaillons qui contient près de 2 900 productions de textes, recueillis du CP au CE2 (Wolfarth *et al.*, 2017).

Au vu de la taille de certains de ces corpus, une exploitation manuelle est très coûteuse et difficilement envisageable. C'est pourquoi, nous cherchons à assister l'exploitation des corpus scolaires qui vise principalement à caractériser linguistiquement les écrits produits en cours d'apprentissage de l'écriture et à analyser les traces visibles de l'évolution de cet apprentissage. À ce titre, opérer un processus de reconnaissance des formes est nécessaire afin de permettre l'extraction du lexique et son analyse, l'analyse des erreurs et de phénomènes linguistiques spécifiques, etc.

Les spécificités de ces corpus et notamment leur grande distance à la norme (pour des exemples voir 2.1) nécessitent des outils spécifiques, inexistants à ce jour. Les outils tels que le *Trameur* (Fleury, 2007) permettent une exploitation textométrique des corpus (y compris scolaires) à condition que ceux-ci soient préalablement annotés. À l'heure actuelle, même si des travaux connexes étudiant l'apport du traitement automatique des langues (TAL) au domaine de l'apprentissage des langues existent (Granger *et al*., 2001 ; Antoniadis *et al*., 2010), peu se sont intéressés au traitement automatique d'écrits scolaires éloignés de la norme, et aucun outil ou méthode ne permet d'opérer les traitements nécessaires (reconnaissance des formes, extraction de lexique, annotation, etc.) de manière automatique ou semi-automatique.

Le travail présenté dans la suite de cet article, qui porte sur la normalisation et la reconnaissance des formes produites dans notre corpus, s'inscrit dans cet objectif et vise à outiller les corpus de textes scolaires. Mais il ne représente qu'une partie du processus que nous souhaitons mettre en œuvre et qui a pour but de proposer une aide à l'exploitation du corpus à des fins de description linguistiques. Il ne s'agit donc que d'une première étape, mais néanmoins nécessaire.

De plus, le travail que nous détaillons ici est un premier travail, réalisé uniquement avec des productions d'apprenants en fin de CP (première année d'apprentissage de l'écrit). Il conviendra donc, par la suite, d'étendre notre étude à l'ensemble du corpus. Cependant, nous pensons que les productions de niveau CP étant particulièrement complexes à traiter en raison du grand nombre d'erreurs présentes dans ces productions, nous pourrons nous inspirer de ces méthodes pour traiter les productions des années suivantes pour lesquelles on note notamment une amélioration de la compétence en segmentation en mots.

Dans la section suivante, nous présenterons des exemples de productions issus de notre corpus et discuterons de la notion de normalisation que nous utiliserons dans la suite de cet article. Puis, dans une troisième partie, nous reviendrons sur divers travaux réalisés autour du traitement de corpus peu normés. Enfin, nous présenterons dans la section 4, l'approche que nous mettons en œuvre pour la reconnaissance des formes dans les corpus d'écrits scolaires, avant de conclure sur les prolongements de notre travail dans la section 5.

# 2 Particularités des corpus scolaires

Dans cette partie, nous présenterons certaines caractéristiques des productions d'apprenants en fin de CP au travers d'exemples et différents choix que nous avons effectués pour notre corpus.

## 2.1 Exemples d'écrits produits en fin de CP

Pour chaque exemple présenté dans cette partie, nous préciserons ses particularités et les types d'écarts à la norme qu'ils contiennent et qui nous intéressent plus particulièrement.

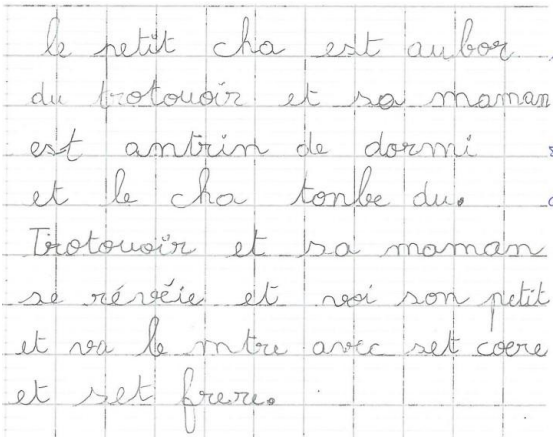



FIGURE 1 : Production de l'élève 96 en fin de CP

Version orthographiée selon la norme[4] : « Le petit chat est au bord du trottoir et sa maman est en train de dormir et le chat tombe du trottoir et sa maman se réveille et voit son petit et va le mettre avec ses sœurs et ses frères. »

Cet exemple illustre une partie des écarts à la norme que produisent les apprenants en fin de CP. On relève en effet la présence de variantes orthographiques, c'est-à-dire de formes graphiques non normées d'un mot, comme les formes *trottoir* et *sœurs*, respectivement orthographiées « trotouoir » et « coere ». On note également des écarts de segmentation en mots, et notamment un cas de sous-ponctuation : « antrin » (*en train*) et des écarts de segmentation en phrase : « tonbe du. Trotouoir » (sur-segmentation). Sur cet exemple, produire une version normée d'un point de vue orthographique est relativement aisé, mais ce n'est pas le cas de toutes les productions : observons l'exemple suivant.

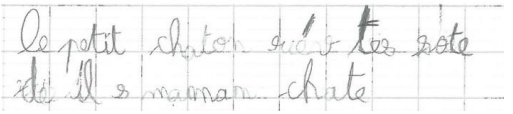


FIGURE 2 : Production de l'élève 2453 en fin de CP

[4] Nous empruntons ici l'appellation « Version orthographiée selon la norme » à C. Boré et M.-L. Elalouf (2017).

Pour cet exemple, la tâche de réécrire la production en normant l'orthographe est davantage une tache d'interprétation que de réécriture. La tâche qui consiste à produire une version orthographiée selon la norme produit donc une version incomplète « le petit chaton se réveille et saute de […] maman chatte ».

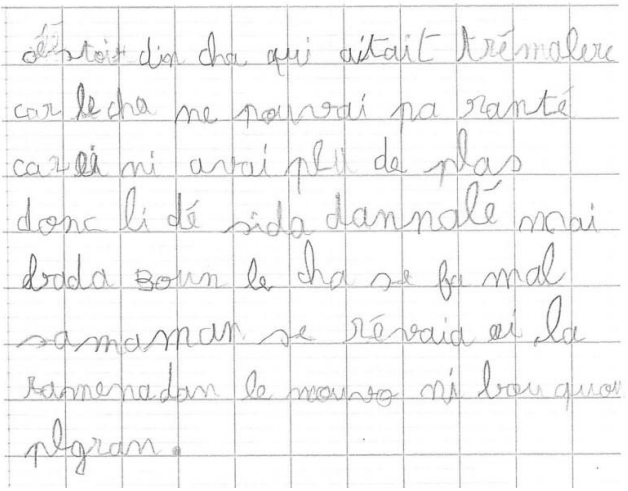


FIGURE 3 : Production de l'élève 2972 en fin de CP

Version orthographiée selon la norme : « C'est l'histoire d'un chat qui était très malheureux car le chat ne pouvait pas rentrer car il n'y avait plus de place donc il décida de s'en aller mais badaboum le chat se fait mal <segmentation/> sa maman se réveilla et le ramena dans le nouveau nid beaucoup plus grand. »

Cet exemple présente également divers types d'écart à la norme. On constate ainsi de nombreuses variantes orthographiques comme « révaia » (*réveilla*), « plas » (*place*) ou encore « malere » (*malheureux*). Cette production présente également des écarts de segmentation en mots, comme dans la séquence *beaucoup plus grand*, produite « bou quou plgran », que ce soit des cas de sous-segmentation : « trémalere » (*très malheureux*), « plgran » (*plus grand*), etc. ou de sur-segmentation : « dé sida » (*décida*). On note également un cas d'absence de marqueur de segmentation en phrases (que ce soit un signe de ponctuation ou un connecteur comme *et* ou *mais*), marqué par la balise <segmentation/> dans la proposition de version orthographiée selon la norme.

Les écarts relevés dans ces productions ne sont que des exemples d'écarts à la norme attestés dans les productions de CP, d'autres auraient pu être mentionnés comme les problèmes d'accord, le non-respect des concordances de temps ou encore des tournures syntaxiques spécifiques mais dans cette première approche des corpus scolaires, nous nous sommes principalement focalisés sur la segmentation en mots et sur les aspects orthographiques.

## 2.2 Intérêt et construction de la normalisation

Nous proposons d'entamer une réflexion en vue d'une aide à l'annotation et à l'exploitation des corpus scolaires, en adoptant une démarche analogue à celle d'O. Kraif et de C. Ponton (2007). Afin

d'envisager l'ensemble des phénomènes décrits dans le paragraphe précédent dans le traitement et l'exploitation des productions d'apprenants, nous proposons une approche s'appuyant à la fois sur une transcription des productions et sur une normalisation de celles-ci. Nous faisons l'hypothèse qu'un alignement forme à forme entre transcriptions et normalisations permettra l'identification des formes produites dans les productions et facilitera donc l'exploitation ultérieure du corpus.

Dans notre corpus, nous disposons donc de trois versions de chaque production :

- Un *scan* de la production manuscrite de l'élève ;
- Une *transcription* tapuscrite réalisée manuellement qui reproduit le contenu linguistique de la production[5]. Précisons que cette étape est réalisée manuellement, les systèmes d'OCR actuels n'étant pas adaptés au type d'écrit de notre corpus ;
- Une *normalisation*.

Si la normalisation utilisée par notre approche n'est pas à construire pour les productions issues d'épreuves de dictées, elle reste à définir pour les productions de textes. Nous pensons que cette approche est généralisable à d'autres formes de productions comme celles issues d'exercices de reformulation ou de copie. Un alignement entre la transcription des productions d'apprenants et leur normalisation devrait alors permettre des allers/retours entre ces deux versions afin de faciliter des interrogations fines de ces corpus, portant par exemple sur le temps des verbes produits par les apprenants, les constructions syntaxiques utilisées. Ce processus permettra également de mesurer les écarts entre formes produites et formes normées afin d'identifier les erreurs les plus fréquentes, ou encore l'évolution de ces erreurs.

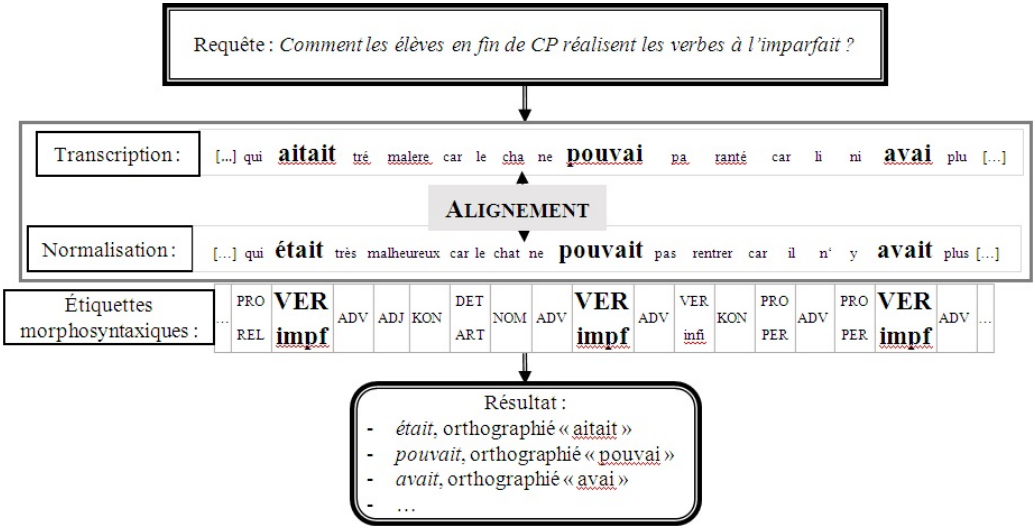



*FIGURE 4 : Exemple de fonctionnement de l'outil et de l'usage de l'algorithme d'alignement à l'aide d'une requête et de la production 2972*

La figure 4 permet d'illustrer un exemple de requête possible grâce à une approche par alignement entre transcriptions et normalisations. Nous donnons ici l'exemple d'une recherche de verbes à l'imparfait. Un étiquetage morphosyntaxique de la normalisation grâce à des outils de TAL déjà existants (TreeTagger dans cet exemple) permettra de retrouver les verbes à l'imparfait sous leur

[5] Pour plus de détails sur cette étape de transcription, voir Wolfarth et al., 2017.

forme normée, tandis que l'algorithme d'alignement permettra de retrouver la forme graphique attestée en corpus de ces verbes. Le résultat d'une telle requête peut être étudié sous différents aspects : usage de l'imparfait dans les productions d'apprenants, réalisation de la finale des verbes en *–ait*, respect ou non de la phonologie dans le cas des erreurs de morphologie verbale, etc.

Afin de réaliser l'alignement, cette transcription est alors comparée à une version normée de la production. Celle-ci doit permettre l'identification des phénomènes que l'on souhaite analyser ; les choix que l'on va effectuer lors de cette étape sont donc essentiels. En effet, tout phénomène non normé ne pourra être retrouvé à l'aide des outils de traitement automatique. Par exemple, si l'on souhaite analyser toutes les formes produites du mot *chat*, il va être nécessaire de normer son orthographe. Si ce cas semble relativement évident, nous avons dû faire des choix bien moins aisés. Considérons l'exemple de la production 1280 (FIGURE 5) : dans la séquence « sa maman réquonsili le bébé chat », il semble que l'enfant a confondu le verbe *réconcilier* et le verbe *réconforter*. Il faudrait normaliser à l'aide du verbe *réconforter*. Mais en procédant à cette substitution, il parait difficile de retrouver le verbe initial dans la transcription puisqu'il n'est pas correctement orthographié et donc non identifiable automatiquement. Dans ce cas, nous avons donc choisi de garder le verbe *réconcilier* et de ne pas le remplacer par le verbe attendu *réconforter*. L'étape de normalisation a fait émerger de nombreuses réflexions similaires.

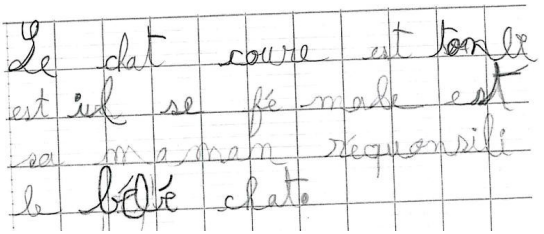


FIGURE 5 : Production de l'élève 2972 en fin de CP

Version orthographiée selon la norme : « Le chat court et tombe et il se fait mal et sa maman réconcilie le bébé chat. »

Dans la plupart des cas, le choix réalisé a été de rester au plus proche de la production de l'apprenant, à l'exception de l'orthographe, de la segmentation en mots et des accords, qui ont été normés. Nous avons également fait le choix de marquer, à l'aide d'une balise, l'absence de segmentation en phrase. La normalisation présente donc une structure syntaxique et lexicale proche de la production initiale.

Les choix que nous présentons ici ont complexes et sont ceux qui ont été adoptés pour calculer les résultats présentés dans la suite de cet article. Il est à noter que ces choix sont amenés à évoluer.

# 3 Travaux connexes

Les travaux de normalisation (selon le sens en vigueur dans le domaine du traitement automatique des langues) et de correction ne sont pas nouveaux en TAL et ont été envisagés dans de nombreux domaines. On peut citer notamment les travaux qui ont cours pour les corpus de SMS, de tweets ou d'écrits issus des réseaux sociaux et des forums en lignes (Kobus *et al*., 2008 ; Beaufort *et al*., 2010 ; Kogkitsidou, 2016). De nombreux travaux ont également vu le jour dans le domaine de l'Apprentissage des Langues Assisté par Ordinateur (ALAO) et de l'apprentissage des langues

secondes (Granger *et al*., 1998 ; Antoniadis *et al.*, 2005 ; Antonsen, 2012). Enfin, on mentionnera également de nombreux travaux dans le domaine de la correction, orthographique le plus souvent, pour scripteurs experts. Mais, à ce jour, peu de travaux concernent l'apprentissage de l'écrit et la normalisation ou la correction de productions d'apprenants de l'écriture.

Dans le cadre de ces différents travaux, diverses approches ont été avancées. Nous présenterons dans cette partie quelques-unes de ces approches. Citons, en premier lieu, les approches par lexiques. Dans ces approches, les formes produites jugées incorrectes ou non normées sont comparées à des formes issues de lexiques selon des indices graphiques, que ce soit à partir des distances d'édition proposées par V. Levenshtein (Kernighan *et al.,* 1990) ou à partir de clés de similarité graphiques (Ndiaye, Vandeventer Faltim, 2004), ou des indices phonologiques (Belrhali, 1995).

Cette approche a été testée sur notre corpus (Wolfarth, 2017) au moyen de comparaisons avec la base lexicale MANULEX (Ortéga, Lété, 2010) basées sur des indices phonologiques, mais la trop grande ambiguïté des résultats de cette approche et la difficulté de la désambiguïsation en raison de la présence de structures syntaxique non encore maitrisées nous poussent à expérimenter une autre méthode.

Certains travaux font également état d'approches par génération d'erreurs (Cohard, 1988 ; Antonsen, 2012). Dans ces approches, de nombreuses formes erronées sont générées à l'aide de règles linguistiques ou d'automates à états finis puis comparées aux formes du corpus en présence. Ces approches sont particulièrement adaptées au traitement des erreurs de morphologie ou aux corpus présentant des règles récurrentes comme les corpus de L2 produits par des locuteurs issus des mêmes L1, mais ne conviennent pas aux productions de CP, comportant un trop grand nombre d'erreurs non régulières.

Enfin, et c'est cette approche qui va nous intéresser dans la suite de cet article, certains travaux envisagent les tâches de correction ou de normalisation comme une tâche d'alignement entre la production attestée en corpus et la production attendue. Cette approche a notamment été abordée dans le cadre de la correction de dictées.

C. Santiago-Oriola (1998) a ainsi conduit une recherche dans une perspective d'aide à la correction de dictées. Dans ce travail, la détection des erreurs est réalisée non pas à l'aide d'un correcteur orthographique mais à l'aide d'un alignement de la dictée et de la production de l'apprenant s'appuyant sur des indices phonologiques et des règles de transformation phonologique.

Plus récemment, R. Beaufort et S. Roekhaut (2011) ce sont appuyés sur un algorithme d'alignement dans la perspective d'automatiser la correction d'exercices de dictées. Il ne s'agissait plus alors de s'appuyer sur des indices phonologiques mais sur une analyse morphosyntaxique automatique de l'original de la dictée. L'alignement est alors opéré grâce à des indices graphiques, issus des opérations d'éditions classiques et de substitutions de séquences graphiques fréquentes au cours de l'apprentissage du français.

Une méthode similaire a été proposée pour un exercice de traduction (Desmet, Hérogue, 2005). Dans ce travail, plusieurs traductions possibles sont générées en guise de version normée. La traduction retenue pour l'alignement est alors la traduction ayant le plus de similitudes avec la réponse de l'apprenant.

Nous proposons ici de tester une approche similaire pour le traitement d'écrits scolaires.

# 4 Alignement par indices phonologiques

Étant donné le choix effectué lors de la normalisation des productions d'apprenants de modifier le moins possible ces productions et le peu de modifications syntaxiques apportées, nous pouvons nous appuyer sur un alignement linéaire des deux versions d'une production. De plus, l'objectif de notre démarche d'alignement s'inscrivant dans un objectif ultérieur d'annotation des formes, l'alignement développé est un alignement de formes, c'est-à-dire de tokens. Toutefois, lorsque cela sera nécessaire à l'annotation, c'est-à-dire lorsque la forme produite et la forme normée ne seront pas identiques, un alignement ultérieur, partiel ou total, de graphèmes, proche de ce que propose C. Santiago-Oriola, pourra être envisagé lors de l'étape d'annotation mais ne concerne pas les propos que nous développons ici.

L'algorithme d'alignement présenté ici a été développé à partir d'un corpus de référence contenant 50 productions de fin de CP, choisies aléatoirement parmi les textes des élèves pour lesquels nous disposions d'un suivi longitudinal et alignées manuellement. À terme, l'aligneur développé devra être adapté à l'ensemble des niveaux de primaires (CP à CM2), il conviendra donc de développer un véritable corpus aligné de référence.

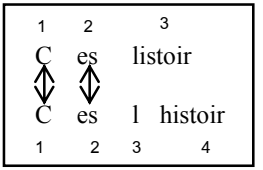
L'aligneur que nous avons développé et que nous présentons par la suite n'est qu'une première version dont l'objectif est double. D'une part, il s'agit de montrer la faisabilité de l'approche et, d'autre part, de disposer d'un premier outil nous permettant d'avancer sur l'exploitation du corpus qui représente le cœur de notre travail. Il est particulièrement conçu pour prendre en considération deux caractéristiques principales des corpus d'apprenants en début d'apprentissage de l'écriture. En effet, il prend en compte les erreurs de segmentations en mots. Cet aspect sera développé dans le paragraphe 4.1. Mais il tient également compte de la dimension phonologique des écrits d'apprenants et s'appuie sur celle-ci, que nous le verrons dans le paragraphe 4.2.

## 4.1 Principe de l'alignement

L'algorithme développé est un algorithme itératif qui considère chaque forme les unes après les autres et ne s'interrompt qu'une fois l'ensemble des formes de chacune des versions examinées. Le principe de l'algorithme est détaillé ci-dessous.

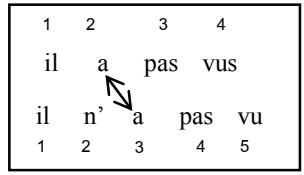
1. Dans le cas le plus courant, l'algorithme procède à une comparaison des formes de rang *n*, selon les mesures exposées au paragraphe suivant (section 4.2). Si les formes sont jugées équivalentes, elles sont alignées, ce sera également le cas pour les étapes suivantes.

```
POUR (n de 0 à longueur(Normalisation)) {
  SI (COMPARER(Transcription[n],Normalisation[n]) == TRUE) {
    ENREGISTRER(Transcription[n],Normalisation[n]) ;
  }
}
```



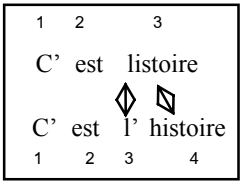

2. Bien que peu de modifications syntaxiques aient été apportées à la normalisation, il se peut que celle-ci comporte tout de même quelques suppressions ou ajouts. L'algorithme permet donc des comparaisons entre la forme produite de rang *n* et la forme normée de rang *n+1*, *n+2* ou *n+3* (cas d'ajout) et inversement (cas de suppression).

```
POUR (n de 0 à longueur(Normalisation) {
  POUR (i de 0 à 3) {
    SI (COMPARER(Transcription[n],Normalisation[n+i]) == TRUE) {
      ENREGISTRER(Transcription[n],Normalisation[n+i]) ;
    }
    SI (COMPARER(Transcription[n+i],Normalisation[n]) == TRUE) {
      ENREGISTRER(Transcription[n+i],Normalisation[n]) ;
    }
  }
```




3. Afin de gérer les erreurs de sous-segmentation, comme « sereveille » (*se réveille*), et de sur-segmentation, comme « dé sida » (*décida*), nous avons également dû implémenter ce que M. Jansche (2001) appelle une *fenêtre glissante* et qui envisage la comparaison de plusieurs formes produites à une forme normée (cas de sur-segmentation) et inversement (cas de sous-segmentation). L'algorithme ainsi développé autorise une fenêtre de longueur 3, considérant qu'une forme produite donnée peut correspondre à trois formes normées et inversement. Cette longueur a été choisie à partir des observations manuelles réalisées au préalable, mais des tests à l'aide de longueurs différentes sont à mener pour déterminer la longueur de fenêtre la plus efficiente.

```
POUR (n de 0 à longueur(Normalisation) {
  SI (COMPARER(Transcription[n], Normalisation[n].Normalisation[n+1]) == TRUE) {
    ENREGISTRER(Transcription[n], Normalisation[n].Normalisation[n+1]) ;
  }
  SI (COMPARER(Transcription[n], Normalisation[n].Normalisation[n+1].Normalisation[n+2]) ==
    TRUE) {
    ENREGISTRER(Transcription[n], Normalisation[n].Normalisation[n+1] .Normalisation[n+2]) ;
  }
  SI (COMPARER(Transcription[n].Transcription[n+1],Normalisation[n]) == TRUE) {
    ENREGISTRER(Transcription[n].Transcription[n+1],Normalisation[n]) ;
  }
  SI (COMPARER(Transcription[n].Transcription[n+1].Transcription[n+2],Normalisation[n]) ==
      TRUE) {
    ENREGISTRER(Transcription[n].Transcription[n+1].Transcription[n+2],Normalisation[n]) ;
  }
}
```




4. Enfin, l'algorithme se déployant de façon linéaire, il permet d'aligner les segments restants, qui ne peuvent être alignés par les mesures explicitées ici.

Si ces étapes permettent la comparaison de plusieurs formes entre elles, chaque comparaison est effectuée selon différentes mesures afin d'améliorer les possibilités d'identifier les équivalences.

## 4.2 Mesures de comparaison

Les travaux de C. Santiago-Oriola nous intéressent particulièrement puisque, partant du constat établi précédemment (Wolfarth, 2017) que seules 25 % des erreurs commises par les apprenants en fin de CP entrainent des modifications d'ordre phonologique, nous choisissons de nous baser sur des indices de type phonologique pour construire un algorithme d'alignement. Comme nous l'avons

mentionné, nous utilisons un algorithme d'alignement linéaire. Celui-ci considère deux formes de rang équivalent, une forme issue de la production de l'apprenant, l'autre issue de la production normée, et les compare selon les modalités suivantes :

1. Comparaison graphique stricte : cette première comparaison permet de repérer les formes correctement orthographiées par les apprenants. Elles sont considérées ainsi lorsqu'il y a correspondance exacte entre forme normée et forme produite;

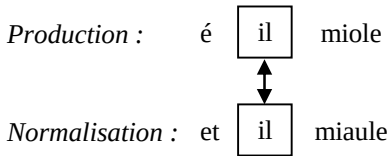



FIGURE 6 : Résultat de la comparaison graphique stricte

2. Comparaison phonologique : il s'agit ici de comparer les représentations phonologiques des formes. Celles-ci sont obtenues à l'aide du module LIA_PHON (Béchet, 2001). Ce module permet de convertir n'importe quelle forme graphique en suite de phonèmes, que cette forme soit ou non correctement orthographiée. La représentation phonologique est produite en suivant les règles de prononciation du français ;

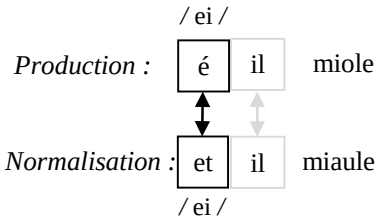



FIGURE 7 : Résultat de la comparaison phonologique stricte

3. Comparaison phonologique avec relâchement de contraintes (ou comparaison archi-phonologique) : nous adoptons ici la notion d'archiphonème, développée par N. Catach (1995). Un archiphonème est un « représentant de l'ensemble des traits phoniques pertinents communs à deux ou plusieurs phonèmes, qui sont par rapport aux autres dans un rapport exclusif » (p.16). Nous regroupons par exemple les phonèmes /ø/, /œ/ et /ə/ sous l'archiphonème /œ/. Ce choix fait suite au constat que les apprenants en fin de CP ne maitrisent pas toujours la distinction entre les phonèmes /ø/, /œ/ et /ə/ ; /e/ et /ɛ/ ; /o/ et /ɔ/ ; /ɛ̃/ et /œ̃/, d'autant plus que les frontières entre ces phonèmes peuvent être différentes d'une région à une autre et qu'ils peuvent être transcrits par un même phonème. Citons pour exemple, les formes *comme* et *comment* où le graphème *o* transcrit à la fois les phonèmes /o/ et /ɔ/. Il s'agit donc ici de comparer les représentations phonologiques des formes produites par LIA_PHON et dont les phonèmes ont été convertis en archiphonèmes

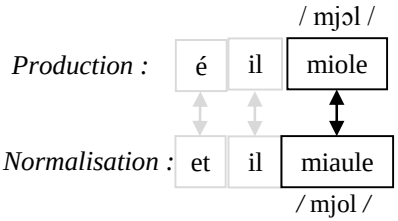



FIGURE 8 : Résultat de la comparaison phonologique avec relâchement de contraintes

Une quatrième voie de comparaison devrait bientôt voir le jour afin de permettre un relâchement des contraintes encore plus étendu grâce à la prise en compte des différentes voies de phonétisation des graphèmes selon leur contexte. Prenons un exemple : la suite de trois lettres *e-s-t* peut être lue /ɛ/ lorsqu'il s'agit du verbe *être* ou /ɛst/ lorsqu'elle est présente dans l'unité lexicale *estimer* ou le point cardinal *est*. C'est cette multiplicité des voies de phonétisation qui amène un enfant de CP qui veut écrire la forme *énorme* à l'orthographier « estnorme ». Dans ce cas, le module LIA_PHON qui suit les règles de lecture du français identifiera la suite phonémique/ɛstnɔrmø/, une prise en compte des différentes possibilités de prononciation d'un graphème ou d'une suite de graphèmes va donc être nécessaire pour retrouver ici la forme normée.

# 5 Résultats et discussion

Cet algorithme d'alignement ayant été élaboré, dans un premier temps, à partir de productions de CP, il a été testé sur un échantillon de 50 productions de fin de CP, soit 1 440 formes. Pour en mesurer l'efficience, nous employons les scores suivants : la précision, le rappel et la F-mesure. Le corpus de référence à partir duquel ces scores sont évalués a été aligné à la main par un seul annotateur.

Ces scores ont été calculés pour chacune des mesures de comparaison présentées au paragraphe 4.2 et sont résumés dans le tableau suivant.

| Mesure | Alignement par comparaison graphique stricte | Alignement par comparaison phonologique | Alignement par comparaison phonologique avec relâchement de contraintes |
|---|---|---|---|
| Précision | 83,3 % | 86,3 % | 86,7 % |
| Rappel | 59,9 % | 72,6 % | 74 % |
| F-mesure | 69,7 % | 78,9 % | 80 % |

TABLEAU 1 : Scores de l'algorithme selon les indices de comparaison utilisés

Les scores de l'algorithme incluant la prise en compte de toutes les règles de prononciation du français n'ont pas encore pu être calculés mais intégrer cette dernière mesure de comparaison devrait permettre d'améliorer quelque peu le score de l'algorithme. Au vu des résultats, il apparait qu'intégrer des indices phonologiques à notre algorithme permet d'améliorer le score de F-mesure de plus de 10 %.

Mentionnons cependant qu'une partie des erreurs provient de la présence de productions telles que les productions 2453 (FIGURE 2) et 208 (FIGURE 10), particulièrement éloignée de la norme, pour lesquelles une interprétation humaine et fort peu aisée. Un traitement automatique de ces productions semble donc proscrit.

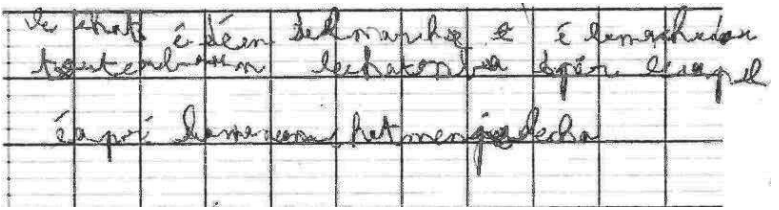


FIGURE 9 : Production de l'élève 208 en fin de CP

Version orthographiée selon la norme : « Le chat est descendu de [la|il][6] marche et la marche […]. Tout à coup boum <segmentation/> le chat tomba et pleure <segmentation/> [le chat|il] appelle et après la mère ramène <segmentation/> le chat mange le chat. »

Cette méthode semble apporter des résultats très intéressants pour notre corpus. Cependant, l'algorithme utilisé dans cet article est un algorithme original, conçu spécifiquement pour le traitement des productions d'apprenant. Il conviendrait donc d'évaluer les approches classiques d'alignement pour les comparer aux résultats produits par cet algorithme.

Évidemment, l'approche par alignement est coûteux puisqu'elle nécessite de produire une normalisation manuellement. La taille des corpus scolaires ne permet cependant pas d'envisager l'emploi d'un apprentissage automatique. Néanmoins, ce premier travail permettra peut-être de produire suffisamment de données d'alignement pour permettre une méthode par apprentissage dans le cas d'un nouveau corpus produit dans un contexte similaire.

Ce travail a également montré la pertinence d'utiliser des indices phonologiques dans le cas de l'apprentissage de l'écrit par des locuteurs natifs. Il pourrait donc être intéressant de mesurer la performance d'un outil spécifique de normalisation inspiré des méthodes de reconnaissance vocale, comme cela a été fait dans certains travaux s'intéressant au traitement des SMS (Kobus *et al.*, 2008). Il conviendrait également d'utiliser d'autres outils de phonétisation afin de comparer leur efficience.

Enfin, cette étude a été réalisée à partir de productions de CP, et nécessitera vraisemblablement des adaptations si l'on s'intéresse à des productions de CE1 et au-delà, au vu des grandes évolutions qui apparaissent dans les premières années d'apprentissage de l'écriture.

# Remerciements

Un grand merci à Claude Ponton et Catherine Brissaud, qui encadrent cette thèse, pour leurs relectures attentives.

[6] Les symboles [$x_1$|$x_2$] marquent le doute entre $x_1$ et $x_2$.

# Références


ANDERSEN H. L., LEBLAY C., AURIAC-SLUSARCZYK E. (2010). Pourquoi travailler sur un corpus commun? Pourquoi travailler de manière pluridisciplinaire?. *Synergies Pays Scandinaves*, (5), 17-30.

ANTONIADIS G., PONTON C., ZAMPA V. (2010). Exxelant et Mirto – Deux exemples d'environnement d'ALAO intégrant des outils TAL. *Multilinguisme et traitement des langues naturelles*. Montréal, Canada : PUQ.

ANTONSEN L. (2012). Improving feedback on L2 misspellings-an FST approach. Actes de *Proceedings of the SLTC 2012 workshop on NLP for CALL*; Lund; 25th October; 2012 (No. 080, pp. 1-10). Linköping University Electronic Press.

BEAUFORT R., ROEKHAUT S., COUGNON L. A., & FAIRON C. (2010). Une approche hybride traduction/correction pour la normalisation des SMS. Actes de *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10)* (pp. 19-23).

BEAUFORT R., ROEKHAUT S. (2011). Automation of dictation exercises. *Computational Linguistics in the Netherlands Journal*, (1), 1-20.

BECHET F. (2001). LIA_PHON - Un système complet de phonétisation de textes, *Traitement Automatique des Langues (T.A.L.)* 42(1), 47-67.

BELRHALI, R. (1995) : Phonétisation automatique d'un lexique général du français : systématique et émergence linguistique. Thèse. Université Stendhal, Grenoble.

BORE C., ELALOUF M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. In Doquet C., David J., Fleury S., *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement*, Corpus, 16 | déc. 2016, 185-214.

CATACH N. (1980, 3e édition, 1995). L'orthographe française : traité théorique et pratique avec des travaux d'application et leurs corrigés (avec la collaboration de Gruaz C. et Duprez D.). Paris : Nathan.

COHARD B. (1988). Logiciel de détection et de correction des erreurs lexicales. Mémoire de master. CNAM.

DAVID J., DOQUET C. (2016). Les écrits d'élèves: un corpus de référence pour le français contemporain. Actes de *SHS Web of Conferences* (Vol. 27, p. 11001). EDP Sciences.

DESMET P., HEROGUEL A. (2005). Les enjeux de la création d'un environnement d'apprentissage électronique axé sur la compréhension orale à l'aide du système auteur IDIOMA-TIC. Actes d *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 8(1), 281–303.

ELALOUF M.-L. (dir) (2005). *Écrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, SCérén, CRDP de Versailles.

FLEURY S. (2007). Le Trameur, Manuel d'utilisation. - http://tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf. 2007

GARCIA-DEBANC C., BONNEMAISON K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans: réussites et difficultés. Actes de *SHS Web of Conferences* (Vol. 8, pp. 961-976). EDP Sciences.

GRANGER S., VANDEVENTER A., HAMEL M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basée sur le TAL. *Traitement automatique des langues* 42(2), 609-621.

KERNIGHAN M. D., CHURCH K. W., GALE W. A. (1990). A spelling correction program based on a noisy channel model. Actes de *Proceedings of the 13th conference on Computational linguistics*-Volume 2 (pp. 205-210). Association for Computational Linguistics.

KOBUS C., YVON F., DAMNATI G. (2008). Normalizing SMS: are two metaphors better than one? Actes de *Proceedings of the 22nd International Conference on Computational Linguistics*(1), Manchester, UK, 441–448.

KOGKITSIDOU E., ANTONIADIS G.. (2016) L'architecture d'un modèle hybride pour la normalisation de SMS. Actes de *la 23e conférence sur le traitement automatique des langues naturelles (TALN'16)* (pp.355-363).

KRAIF O., PONTON C. (2007). Du bruit, du silence et des ambiguïtés: que faire du TAL pour l'apprentissage des langues. Actes de *la 14e conférence sur le traitement automatique des langues naturelles (TALN'07).*

NDIAYE M. ET VANDEVENTER FALTIN A. (2004). Correcteur orthographique adapté à l'apprentissage du français. BULAG, 29:117–134.

ORTÉGA É., LÉTÉ B. (2010). « eManulex: Electronic version of Manulex and Manulex-infra databases », http://www.manulex.org

SANTIAGO ORIOLA C. (1998). Système vocal interactif pour l'apprentissage des langues. La synthèse de la parole au service de la dictée. Thèse. Toulouse III.

WOLFARTH C., PONTON C., BRISSAUD C. (2016) Du TAL dans les écrits scolaires: premières approches. Actes de *RECITAL'16 - rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*.

WOLFARTH C., PONTON C., TOTEREAU C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique. *Corpus 16*, janv. 2017, 185-214.

# Générer une grammaire d'arbres adjoints pour l'arabe à partir d'une méta-grammaire

Cherifa Ben Khelil 1, 2
(1) LIFO, Bâtiment IIIA 6 Rue Léonard de Vinci, F-45067 Orléans, France
(2) RIADI, Campus Universitaire de la Manouba, 2010 Manouba, Tunisie
cherifa.ben-khelil@etu.univ-orleans.fr cherifa.bk@gmail.com

RÉSUMÉ
La rareté des ressources numériques pour la langue arabe, telles que les grammaires et corpus, rend son traitement plus difficile que les autres langues naturelles. A ce jour il n'existe pas une grammaire formelle à large couverture de l'arabe. Dans ce papier, nous présentons une nouvelle approche qui facilite la description de l'arabe avec le formalisme des grammaires d'arbres adjoints en utilisant une méta-grammaire. Nous exposons les premiers résultats de notre grammaire ainsi que les problèmes rencontrés pour son évaluation.

ABSTRACT
**Generate a tree adjoining grammar for arabic from a meta-grammar**

The scarcity of digital resources for arabic language, such as grammars and corpus, makes its processing more difficult than other natural languages. To date there is not a wide-coverage grammar of the Arabic langage. In this paper, we present a new approach in order to facilitate the description of Arabic with Tree Adjoining Grammar using a meta-grammar. We present our first results as well as problems encountered in the evaluation of the grammar..

MOTS-CLÉS : Arabe, grammaire d'arbres adjoints, méta-grammaire, évaluation.

KEYWORDS: Arabic, Tree Adjoining Grammar, meta-grammar, evaluation.

## 1 Introduction

En la linguistique et l'informatique, une grammaire formelle constitue un moyen de représentation permettant de définir une syntaxe ou plus largement un langage formel. Elle est utilisée dans plusieurs applications dans le domaine du Traitement Automatique du Langage Naturel (TALN) (systèmes de traduction automatiques, dialogue homme-machine, etc.).
Nous nous sommes intéressés à la langue arabe qui est une langue sémitique. L'arabe présente beaucoup de spécificités à différents niveaux : phonologique, morphologique, syntaxique et aussi sémantique. Bien que plusieurs travaux de recherche aient abordé ces problématiques, les ressources numériques utiles pour le traitement de l'arabe demeurent relativement rares. En effet, la majorité de ces ressources libres traite de la morphologie (Xerox Arabic Morphological Analysis and Generation [1], ElixirFM [2], etc.) mais peu d'entre elles sont disponibles pour la syntaxe. Nous pouvons citer la grammaire générative transformationnelle (Alkhuli, 1979; Kebbe, 2000), d'autres modèles linguistiques

1. https ://open.xerox.com/Services/arabic-morphology
2. https ://github.com/otakar-smrz/elixir-fm

tels que la grammaire lexicale fonctionnelle (LFG) (Fehri, 1981) et la grammaire syntagmatique guidée par les têtes (HPSG) (Mutawa *et al.*, 2008; Haddar *et al.*, 2010) . Cependant, ces grammaires ont une couverture limitée et à ce jour il n'existe pas un analyseur syntaxique à large couverture de l'arabe. Outre les grammaires, les corpus formés de phrases analysées et stockées sous forme de banques d'arbres (Treebanks) constituent un autre type de ressources syntaxiques mais ils ne sont pas en accès libre.

C'est dans ce contexte que s'inscrit notre travail de recherche qui vise à élaborer une grammaire formelle pour l'utiliser dans les applications du traitement automatique de la langue arabe. Notre choix s'est porté sur le formalisme des grammaires d'arbres adjoints (TAG)(Joshi *et al.*, 1975). Ce choix est motivé par le pouvoir de représentation des TAG (les structures simples, complexes, combinatoires, partagées, etc...) et leur capacité de traiter certains phénomènes tels que les enchâssements. La grammaire que nous proposons est produite semi automatiquement grâce au le langage de description méta-grammatical XMG (eXtensible MetaGrammar) (Crabbé *et al.*, 2013).

Cet article est organisé de la manière suivante. Dans la section 2, nous présentons le formalisme TAG et son application à l'arabe. Dans la section 3, nous exposons notre approche qui décrit l'arabe au moyen d'une méta-grammaire. Finalement, dans la section 4, nous discuterons des problèmes rencontrés lors de l'évaluation de la qualité de la grammaire.

# 2 Représentation de la syntaxe de l'arabe au moyen de TAG

## 2.1 Présentation du formalisme TAG

La grammaire d'arbres adjoints (Joshi *et al.*, 1975) est un formalisme syntaxique qui permet de rendre compte des liens entre les constituants de la phrase. Il offre un système de réécriture d'arbres dont les unités sont des arbres élémentaires. Ceux-ci sont définis par un ensemble d'arbres initiaux et d'arbres auxiliaires.

- Un arbre initial est un arbre fini, dont les nœuds feuilles sont soit des symboles terminaux, soit des symboles non terminaux. Les symboles non terminaux sont appelés nœuds de substitution et sont marqués par le symbole ($\downarrow$).
- Un arbre auxiliaire a en outre un nœud feuille étiqueté par un symbole non terminal appelé "nœud pied" et il est marqué par le symbole (*). Le nœud pied et la racine de l'arbre auxiliaire sont nécessairement de même catégorie.

Les deux opérations de réécriture d'arbres autorisées par TAG sont : l'adjonction et la substitution. A l'issue de ces opérations de réécriture, on obtient un nouvel arbre appelé arbre dérivé.

L'opération de substitution (Figure 1) permet d'insérer un arbre $\alpha$ à la frontière d'un arbre dérivé initial contenant un nœud de substitution $\beta$. La substitution est autorisée uniquement si le nœud de substitution et le nœud racine, respectivement de $\beta$ et de $\alpha$ sont étiquetés par un symbole identique.

L'opération d'adjonction (Figure 2) permet, plutôt, d'insérer l'arbre $\gamma$ dans l'arbre $\beta$ sur un nœud interne X. Le nœud X situé dans $\beta$ est remplacé par la racine de $\gamma$. L'adjonction est autorisée si la catégorie du nœud X est identique à la catégorie du nœud racine de $\gamma$.

TAG est légèrement sensible au contexte (mildly contextsensitive) (Weir, 1988). C'est un formalisme plus puissant que les grammaires hors contextes [3], mais strictement incluse dans la classe des

3. C'est une grammaire de réécriture dont les parties gauches des règles contiennent un unique non-terminal, donc sa dérivation ne dépend d'aucun contexte (X $\longrightarrow$ w)

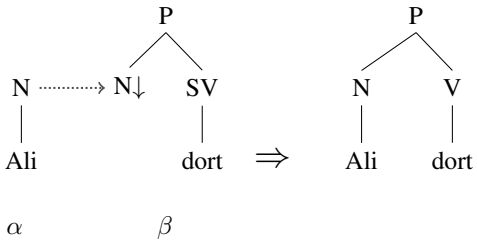



FIGURE 1 – Substitution de l'arbre $\alpha$ dans l'arbre $\beta$

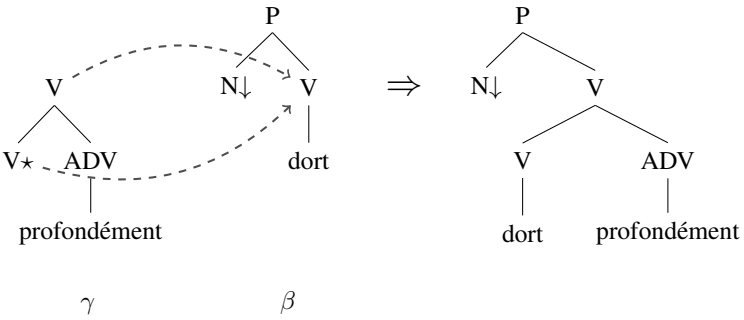



FIGURE 2 – Adjonction de l'arbre $\gamma$ dans l'arbre $\beta$

grammaires contextuelles [4]. Elle définit un domaine de localité étendu étant donné que la profondeur des arbres élémentaires est variable, au contraire des règles de réécriture hors contexte dont la profondeur est égale à 1. De ce fait, elle possède un fort pouvoir génératif, qui englobe les dépendances à longue distance et certaines dépendances croisées ainsi qu'une factorisation des composantes grammaticales récursives. De plus, d'un point de vue traitement, TAG reste analysable en un temps polynomial O ($n^6$). Nous ne pouvons pas affirmer que ce formalisme est incontestablement le meilleur pour représenter l'arabe. Néanmoins, ses caractéristiques ont permis de représenter des phénomènes de l'arabe tels que l'enchâssement.

## 2.2 Capacité à gérer l'enchâssement

La représentation de ce phénomène est possible avec TAG grâce à l'opération d'adjonction.Cette opération permet d'intercaler une structure complète dans une autre structure rendant ainsi la représentation des dépendances enchâssées très naturelle. L'adjonction d'un arbre auxiliaire à lui-même ou à un autre arbre élémentaire met en valeur la récursivité de la langue naturelle.

Considérons l'exemple (Figure 3) du syntagme nominal de liaison (مركب موصولي) الذي حقق النجاح

Ce type de grammaire permet de définir les langages algébriques qui sont reconnaissables par un automate à pile non déterministe.

4. Les règles d'une grammaire contextuelle sont restreintes en obligeant le membre droit à être au moins aussi long que le membre de gauche ( uXv $\longrightarrow$ uwv)
Le langage engendré par ce type de grammaire est récursif et reconnaissable par un automate borné linéairement.

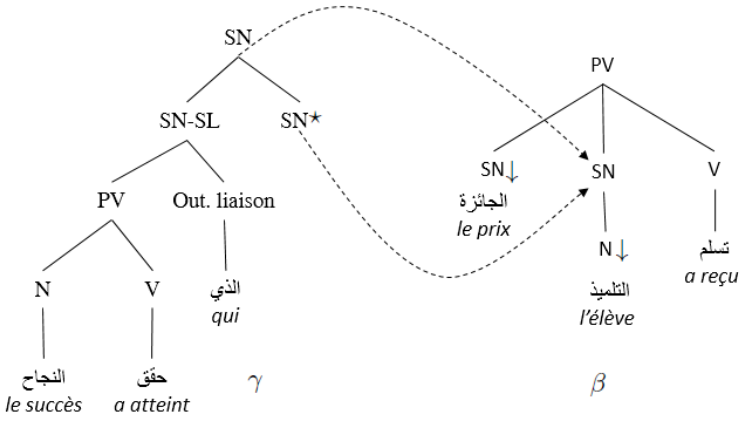



FIGURE 3 – Exemple représentatif de la manipulation des structures enchâssées en TAG

( qui a atteint le succès ), représenté par l'arbre $\gamma$, il peut être enchâssé dans une phrase principale تسلم التلميذ الجائزة (l'élève a reçu le prix) dont la structure est représentée par l'arbre $\beta$, et donner ainsi naissance à une nouvelle phrase plus complexe qu'est : تسلم التلميذ الذي حقق النجاح الجائزة (l'élève qui a atteint le succès a reçu le prix).

## 2.3 ArabTAG V1.0 (Arabic Tree Adjoining Grammar)

A notre connaissance, il existe deux travaux qui ont construit une TAG pour l'arabe : une TAG par extraction d'arbres élémentaires à partir du corpus Penn Arabic TreeBank (Habash & Rambow, 2004) et ArabTAG (Ben Fraj, 2011). Notre travail de thèse prend ses origines de cette dernière.
ArabTAG (Arabic Tree Adjoining Grammar) hérite de tous les fondements de base de TAG.
Elle décrit différents composants syntaxiques de différents niveaux : phrases, syntagmes et mots ainsi que les différentes informations qui leur sont relatives (morphologiques et syntaxiques).
Dans le but de limiter le nombre d'arbres élémentaires, seules les structures syntaxiques, formées des éléments essentiels de la phrase, sont représentées telles que les phrases verbales (Verbe, sujet et complément d'objet), phrases nominales et les syntagmes nominaux.
De plus, pour la bonne gestion des compositions syntaxiques, (Ben Fraj, 2010) a défini différents types d'informations (morphologiques, syntaxiques) au sein d'une structure dite « traits d'unification » (Joshi *et al.*, 1975) associés à chaque nœud de l'arbre élémentaire. En plus de ces traits d'unification, elle a défini un ensemble de « traits d'instanciation ». Ces derniers sont attribués seulement aux nœuds ancres et définissent les informations morphosyntaxiques nécessaires lors de la lexicalisation de la grammaire.
ArabTAG est partiellement lexicalisée puisqu'elle contient un ensemble d'arbres élémentaires lexicalisés réservés pour représenter les contextes possibles des mots outils jouant le rôle de « modificateurs » des phrases et/ou des syntagmes.
La première version d'ArabTAG est composée de différentes structures syntaxiques réparties entre

phrases et syntagmes. Elle contient 380 arbres élémentaires différents, répartis entre quatre grandes familles : phrases verbales, phrases nominales, syntagmes nominaux et syntagmes prépositionnels. Elle couvre quelques structures elliptiques, d'autres anaphoriques et aussi des structures renfermant des subordonnées. Elle prend en considération la variation de l'ordre des éléments au sein des phrases et aussi le phénomène d'agglutination.
Nous avons étudié cette première version de la grammaire et nous avons relevé certaines limites pouvant se résumer comme suit :

— une couverture minimale : les structures syntaxiques possibles ne sont pas toutes décrites. Elle ne représente pas, par exemple, les structures enrichies avec des compléments (circonstanciel de temps, de lieu, etc).
— la représentation des formes agglutinantes dans les structures syntaxiques n'est pas bien prise en compte. Ces formes peuvent jouer des rôles dans la phrase et doivent être mises en relief pour améliorer la couverture du modèle grammatical développé.
— ArabTAG V1.0 met l'accent sur les relations syntaxiques sans s'intéresser aux informations sémantiques, bien que la sémantique, tout comme la morphologie, possède une influence directe sur la syntaxe. En effet, l'interprétation syntaxique ne peut être complète que si l'on fait intervenir des informations sémantiques.
— ArabTAG V1.0 n'est pas organisée en des structures factorisées hiérarchiquement. Elle est composée d'un ensemble d'arbres élémentaires sans qu'ils soient reliés entre eux.Dans le but de faciliter la maintenance et l'extension de la grammaire, il est primordial de structurer la grammaire en faisant intervenir divers phénomènes tels que l'héritage des structures ou la hiérarchie des patrons d'arbres.

Par conséquent, nous proposons une nouvelle version ArabTAG V2.0 afin d'intégrer davantage les aspects syntaxiques et bénéficier des avantages représentatifs de TAG.

# 3 Description de l'arabe au moyen d'une méta-grammaire

La nouvelle version de la grammaire : ArabTAG V2.0 (Ben Khelil *et al.*, 2016) est réécrite en utilisant le formalisme XMG (Crabbé *et al.*, 2013). Ce dernier présente des caractéristiques particulièrement pertinentes pour la description des arbres élémentaires pour la langue arabe :

— il est très expressif, ce qui permet de définir des descriptions factorisées de la grammaire (dans notre cas, il a été utilisé pour traiter l'ordre des mots semi-libre).
— il est particulièrement adapté à la description des grammaires d'arbres et a été utilisé pour développer plusieurs grammaires TAG électroniques comme par exemple le français [5], l'anglais [6] et l'allemand [7] ;
— il est extensible et peut être configuré pour décrire différents niveaux de langage, comme la sémantique ou la morphologie.

A l'aide de ce formalisme, nous avons généré notre TAG semi-automatique pour l'arabe à partir d'une description réduite des règles de la grammaire. Cette description compacte de l'information grammaticale correspond à une méta-grammaire. Elle capture les généralisations linguistiques apparaissant parmi les arbres de la grammaire et permet ensuite de générer les arbres TAGs correspondants.
ArabTAG V2 est organisée en des structures factorisées hiérarchiquement grâce au mécanisme d'héritage de XMG. Ce dernier permet d'automatiser la combinaison des fragments d'arbres, que nous

5. https ://sourcesup.renater.fr/xmg/frenchmetagrammar/index.html
6. http ://homepages.inf.ed.ac.uk/s0896251/XMG-basedXTAG/titlepage.html
7. http ://www.sfs.uni-tuebingen.de/emmy/res.html

avons définis, en attribuant une couleur (B, W et R pour noir, blanc et rouge respectivement) aux nœuds. La couleur rajoute une contrainte de bonne formation des arbres engendrés : Un nœud coloré en noir peut être unifié avec 0, 1 ou plusieurs nœuds blancs et produit ainsi un nœud noir. Un nœud blanc doit être unifié avec un noir produisant un nœud noir. Finalement, un nœud rouge ne peut pas être fusionné avec un autre nœud.
Nous avons commencé par définir une classe EpineVerbe(C) [8] qui contribue à un fragment d'arbre pour l'épine verbale (classe MorphActive). Nous avons rajouté des points d'adjonction appropriés pour les adverbes [9]. Ces derniers peuvent être librement intercalés entre arguments.

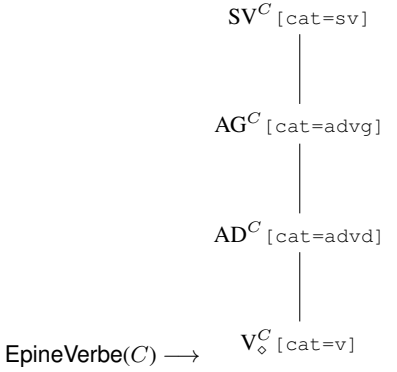



FIGURE 4 – Description de la classe EpineVerbe

La classe MorphActive instancie EpineVerbe. La valeur de la couleur passée en paramètre est noir :

MorphActive $\longrightarrow$ EpineVerbe(B)

La classe SujetCanon représente le sujet du verbe (cas nominatif) :

SujetCanon $\longrightarrow$ [AD] $\Leftarrow$ EpineArg() $\wedge$ SN$^{R}_{\downarrow}$ [cat=sn,cas=nom] $\wedge$ AD $\triangleleft$ SN

Quant aux compléments d'objets directs et indirects ils sont définis comme suit : Le complément d'objet direct est exprimé par un syntagme nominal [cat=sn] dont le cas est accusatif [cas=acc]. Ce complément d'objet peut se placer avant le verbe. Dans ce cas, une anaphore (référant au complément d'objet) doit s'attacher à la fin du verbe. Cette contrainte est exprimée avec le trait [oclit=+] [10] :

ObjetCanonSN $\longrightarrow$ [AD, V] $\Leftarrow$ EpineArg() $\wedge$ SN$^{R}_{\downarrow}$ [cat=sn, cas=acc] $\wedge$ AD $\triangleleft$ SN
$\wedge$ ((V [oclit=−] $\wedge$ V $\prec^{+}$ SN) $\vee$ (V [oclit=+] $\wedge$ SN $\prec^{+}$ V))

La classe suivante indique que le complément d'objet peut être un enclitique au verbe :

ObjetCanonClit $\longrightarrow$ [V] $\Leftarrow$ EpineVerbe(W) $\wedge$V [oclit=+]

Quant au complément d'objet indirect, il est exprimé par un syntagme prépositionnel :

ObjetIndCanon $\longrightarrow$ [AD] $\Leftarrow$ EpineArg() $\wedge$ SP$^{B}$ [cat=sp] $\wedge$ AD

P$^{R}_{\diamond}$ SN$^{R}_{\downarrow}$ [cat=sn, cas=gen]

$\triangleleft$ SP

8. EpineVerbe (Figure 4) est paramétré par une couleur C
9. AG pour adverbe à gauche et AD pour adverbe à droite
10. Trait qui indique la présence d'un enclitique liée au nœud en question.

Au final, nous avons utilisé la transitivité du verbe comme un critère fondamental pour l'héritage et nous avons combiné [11] ces fragments afin d'obtenir les trois familles de verbes (intransitif, transitif, ditransitif). Chacune de ces classes capture les différentes réalisations syntaxiques possibles entre les différentes structures de la phrase. Elles sont organisées comme suit :
Intransitive $\rightarrow$ MorphActive $\wedge$ (SujetCanon $\vee$ Ellipse)
La classe Morphactive définit le fragment d'arbre élémentaire qui constitue l'épine verbale de la phrase. Suite à une opération de conjonction entre cette classe et les classes du sujet (SujetCanon ou Ellipse du sujet), nous obtenons la classe Intransitive.
Transitive $\rightarrow$ Intransitive $\wedge$ ObjetCanon[Objet1] [12]
La classe Transitive est obtenue en combinant la classe Intransitive est la classe ObjetCanon. Cette dernière représente une disjonction entre le complément d'objet direct, indirect et clitique.
DiTransitive $\rightarrow$ Transitive $\wedge$ ObjetCanon[Objet2] [13]
Finalement, nous obtenons la classe DiTransitive en combinant Transitive avec les compléments d'objets seconds.
L'utilisation de XMG nous a permis de gérer facilement le problème d'ordre semi-libre des mots. Pour ce faire, nous avons évité d'imposer une contrainte de précédence entre les arguments dont le changement d'ordre n'affecte pas la cohérence de la phrase.
Jusqu'à présent, nous avons généré 315 arbres à partir d'une description faite de 28 classes (soit 28 fragments d'arbres ou règles de combinaison). Cette version de la grammaire couvre les structures déjà couvertes par la première version d'ArabTAG à savoir les phrases verbales (forme active et passive), phrases nominales, syntagmes nominaux et syntagmes prépositionnels. La couverture de la grammaire a été également étendue en ajoutant des arbres élémentaires pour la représentation des compléments tels que les compléments circonstanciels de temps, compléments circonstanciels de lieu et les adverbes.
Nous avons mis en place un environnement de développement (script en langage Python) afin de vérifier la couverture grammaticale d'ArabTAG V2.0. En plus de notre grammaire, nous avons défini des lexiques syntaxiques et morphologiques en suivant l'architecture en 3 couches du projet (XTAG Research, 2001). En effet, le système XTAG se compose de trois sous-modules :

- Une base de schèmes qui sont classés en familles d'arbres élémentaires.
- Une base de lemmes où chaque lemme est associé une (ou plusieurs) famille(s) d'arbres et à un prédicat sémantique.
- Une base morphologique dans laquelle chaque forme fléchie est associée à un lemme et l'information morphosyntaxique appropriée.

Le but de ce test (Figure 5) est d'évaluer à la fois la sous-génération et la surgénération. A chaque nouveau phénomène syntaxique inclus dans ArabTAG V2.0, le corpus de test est enrichi manuellement avec des phrases grammaticales et non grammaticales (associées aux nombres d'analyses syntagmatiques attendues). L'analyseur TuLiPA (Parmentier *et al.*, 2008) est ensuite exécuté sur le corpus de test pour vérifier la qualité de la grammaire. Les résultats des analyses nous aident à corriger les erreurs potentielles dans la description de notre grammaire et nous garantit la cohérence des structures TAG lors de son extension.

11. XMG offre des opérations de combinaisons par disjonction et conjonction des fragments
12. ObjetCanon[Objet1] $\longrightarrow$ ObjetCanonSN[Objet1] $\vee$ ObjetCanonClit[Objet1] $\vee$ ObjetIndCanon[Objet1]
13. ObjetCanon[Objet2] $\longrightarrow$ ObjetCanonSN[Objet2] $\vee$ ObjetCanonClit[Objet2] $\vee$ ObjetIndCanon[Objet2]

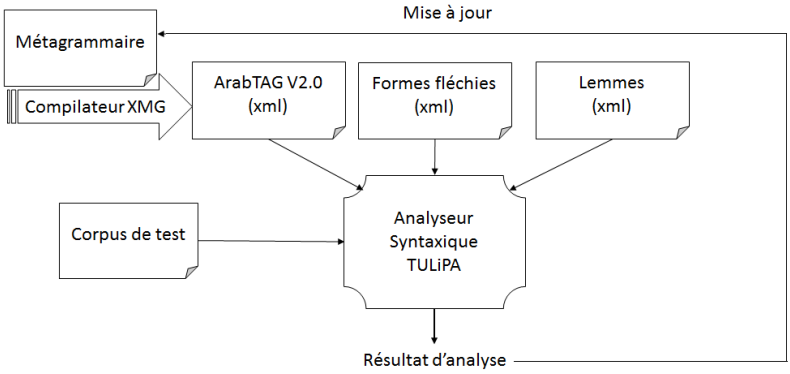



FIGURE 5 – Architecture du module de validation d'ArabTAG v2.0

# 4 Évaluer la qualité de la grammaire

L'étape d'évaluation est primordiale et doit prendre en compte un certain nombre de critères. Il est indispensable de s'interroger non seulement sur les ressources nécessaires à une telle évaluation mais aussi la procédure à adopter et quels critères d'évaluation utiliser.

## 4.1 Critères d'évaluation

Tout d'abord, et avant de s'engager dans cette étape, nous avons fait le point sur ce que notre grammaire doit satisfaire. En effet, la grammaire doit pouvoir :

- Distinguer les phrases valides telles qu'elles sont décrites dans l'arabe standard moderne (livres scolaires, romans arabes, etc.)
- Distinguer les phrases agrammaticales
- Analyser un ensemble assez grand de phrases qui couvrent les phénomènes syntaxiques importants de l'arabe.

Nous avons commencé par définir les critères de l'évaluation. Les métriques classiques utilisées sont la précision et le rappel. La précision est le nombre de phrases grammaticales bien analysées par rapport au nombre total de phrases analysées. Quant au rappel, il est défini par le nombre de phrases grammaticales bien analysées par rapport au nombre total de phrase grammaticales présentes dans le corpus de test. Ces deux métriques sont combinées dans une mesure nommée F-mesure (ou encore le f-score). La quantification de ces critères se fait donc sur la base d'un corpus annoté, typiquement un corpus arboré.

## 4.2 Protocole d'évaluation

Les outils dont nous disposons (présentés dans la section précédente) ne peuvent pas supporter un lexique de grande taille, ce qui ne nous permet pas de conclure sur la qualité de la grammaire. En effet, nous avons défini le lexique et le corpus de test manuellement. Le corpus de test est constitué de 120 phrases et syntagmes dont 32 agrammaticales. Nous voulons donc évaluer notre grammaire en utilisant un corpus arboré de grande taille. Le jeu de test sera constitué d'un ensemble de phrases et de syntagmes. L'étape suivante consistera à extraire automatiquement à partir du corpus les éléments nécessaires pour constituer un lexique compatible avec notre grammaire.

## 4.3 Ressources et Corpus d'évaluation

La difficulté majeure que nous avons rencontrée dans cette étape concerne les ressources nécessaires à l'évaluation. En effet, la disponibilité de ces ressources est contrastée selon les langues. Pour l'arabe standard, un ensemble de trois corpus arborés a été créé : Penn Arabic TreeBank [14](PATB) (Maamouri *et al.*, 2004),Prague Arabic Dependency Treebank [15] (PADT) (Hajič *et al.*, 2004) et Columbia Arabic Treebank [16] (CATiB) (Habash & Ryan, 2009). Ces corpus n'adoptent pas le même format d'annotation. PATB repose sur les représentations à base des structures de constituants [17] tandis que PADT et CATiB suivent, plutôt, un format à base des structures de dépendances [18].Les textes inclus dans ces corpus sont des textes journalistiques. Or nous voulons exploiter des textes littéraires renfermant des structures plus riches et plus représentatives de l'arabe et de ses phénomènes syntaxiques. En plus, tous ces corpus ne sont pas des ressources libres et leurs coûts sont élevés. Il nous est difficile de les utiliser dans notre évaluation.

## 4.4 Proposition pour ArabTAG V2.0

Face à la contrainte de disponibilité des ressources pour l'évaluation, nous avons décidé d'utiliser un ensemble de textes étiquetés. Ces textes, initialement bruts [19], ont subi un ensemble de traitements séquentiels [20] qui ont permis de les transformer en des phrases étiquetées par les arbres syntaxiques corrects qui leur correspondent (Ben Fraj *et al.*, 2009). Ce corpus a été construit et utilisé dans la procédure d'analyse syntaxique à base d'apprentissage pour la langue arabe (Ben Fraj, 2010).
Ces textes étiquetés seront le point de départ pour construite un corpus comportant des phénomènes précis de l'arabe. Nous pouvons citer l'un des corpus de phénomènes syntaxiques de l'anglais Phenomenal Corpus (Letcher & Baldwin, 2013) qui a été annoté manuellement. Chaque phrase de notre corpus sera associée à un phénomène syntaxique de l'arabe que nous voulons tester. L'étape suivante consistera à extraire automatiquement, à partir du corpus étiqueté, les éléments nécessaires à notre lexique, à savoir la liste des formes fléchies avec leurs traits ainsi que la liste des lemmes

14. https ://catalog.ldc.upenn.edu/LDC2003T06
15. https ://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html
16. http ://www1.ccls.columbia.edu/ CATiB/Home.html
17. Les données sont étiquetées à trois niveaux : morphologique, syntaxique et grammatical.
18. Ces corpus utilisent un format de dépendances simple étiqueté par les relations fonctionnelles des différents composants syntaxiques.
19. Les textes sources ont été collectés du livre de la langue arabe de la 8ème année de base de l'enseignement tunisien (2007-2008).
20. une analyse morphosyntaxique, un étiquetage grammatical, une segmentation en phrases et finalement un étiquetage syntaxique. Certains de ces pré-traitements ont été suivis d'une vérification manuelle.

correspondants. Les traits seront automatiquement (ou semi-automatiquement) attribués lors de l'extraction grâce à un procédé de correspondance entre les jeux d'étiquettes du corpus et ceux utilisés dans notre grammaire. Par conséquent, ce lexique est spécifique à notre corpus. Néanmoins, nous envisageons son enrichissement en utilisant un dictionnaire arabe. Le corpus sera agrandi avec des nouveaux cas plus complexes ainsi que des phrases agrammaticales.

# 5 Conclusion

L'analyse syntaxique de la langue arabe est une tâche difficile à entreprendre. Cela explique la rareté des ressources numériques disponibles pour le traitement automatique de cette langue ; à savoir les grammaires et les corpus. Dans cet article, nous avons présenté une nouvelle approche visant à construire une grammaire d'arbres adjoints pour représenter la syntaxe de l'arabe. Cette grammaire est réécrite en utilisant un langage de description méta grammatical XMG. Ce langage extensible fournit une représentation compacte de l'information grammaticale et offre un mécanisme pour combiner les fragments élémentaires d'information. Il permet ainsi le partage des informations entre les structures grammaticales et peut être configuré pour décrire différents niveaux du langage.
La grammaire couvre les phrases verbales (forme active et passive), phrases nominales, syntagmes nominaux et syntagmes prépositionnels. Elle traite la variation des positions des éléments au sein des composants syntaxiques, les compléments supplémentaires et les formes agglutinées.
De plus, elle utilise les traits d'unification. Ces traits présentent des informations morphologiques, syntaxiques et syntaxico-sémantiques supplémentaires qui sont associées à un mot ou à un syntagme.
Actuellement, nous voulons évaluer notre grammaire. Cependant l'indisponibilité des ressources nécessaires pour une telle évaluation nous a contraints à utiliser un petit corpus de test qui sera agrandi avec d'autres exemples complexes de phrases.
Nous envisageons d'améliorer la couverture d'ArabTAG V2.0 en intégrant les dimensions morphologiques et sémantiques. Pour ce faire, nous étudions l'utilisation éventuelle de XMG2 (Petitjean, 2014) qui étend XMG en incluant un compilateur méta-grammatical. Ceci rend possible la description des nouvelles dimensions, qui sont respectivement la description de la morphologie en utilisant les champs topologiques et celle de la sémantique en utilisant les frames.

# Remerciements

Nous remercions les relecteurs pour leurs commentaires pertinents qui ont permis d'améliorer la qualité de ce papier.

# Références

ALKHULI M. (1979). *A contrastive transformational grammar : Arabic and English*. Leiden E. J. Brill.

BEN FRAJ F. (2010). *Un analyseur syntaxique pour les textes en langue arabe à base d'un apprentissage à partir des patrons d'arbres syntaxiques*. PhD thesis, ENSI La Manouba, Tunisia.

BEN FRAJ F. (2011). Construction d'une grammaire d'arbres adjoints pour la langue arabe. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France : Association pour le Traitement Automatique des Langues.

BEN FRAJ F., BEN OTHMANE ZRIBI C. & BEN AHMED M. (2009). A semi-automatic tag syntactic tagging tool for constructing an arabic treebank. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, Mrągowo , Pologne.

BEN KHELIL C., DUCHIER D., PARMENTIER Y., ZRIBI C. & BEN FRAJ F. (2016). ArabTAG : from a Handcrafted to a Semi-automatically Generated TAG. In *TAG+12 : 12th International Workshop on Tree-Adjoining Grammars and Related Formalisms*, Düsseldorf, Germany.

CRABBÉ B., DUCHIER D., GARDENT C., LE ROUX J. & PARMENTIER Y. (2013). XMG : eXtensible MetaGrammar. *Computational Linguistics*, **39**(3), 591–629.

FEHRI A. (1981). *Complémentation et anaphore en arabe moderne : une approche lexicale fonctionnele*. Doctoral dissertation.

HABASH N. & RAMBOW O. (2004). Extracting a tree adjoining grammar from the penn arabic treebank. *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*, p. 277–284.

HABASH N. & RYAN M. R. (2009). Catib : The columbia arabic treebank. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009 ACL and AFNLP, Suntec, Singapore*.

HADDAR K., BOUKEDI S. & ZALILA I. (2010). Construction of an hpsg grammar for the arabic language and its specification in tdl. *International Journal on Information and Communication Technologies*, **3**(3).

HAJIČ J., SMRŽ O., ZEMÁNEK P., ŠNAIDAUF J. & BEŠKA E. (2004). Prague arabic dependency treebank : Development in data and tools. *In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools,Cairo, Egypt*.

JOSHI A., LEVY L. & TAKAHASHI M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences*, **10**(1), 136 – 163.

KEBBE M. (2000). *Transformational Grammar of Modern Literary Arabic*. Routledge.

LETCHER N. & BALDWIN T. (2013). Constructing a phenomenal corpus : Towards detecting linguistic phenomena in precision grammars. In *Workshop on High-level Methodologies for Grammar Engineering @ ESSLLI 2013*, p. 367–376, Düsseldorf, Germany.

MAAMOURI M., BIES A., BUCKWALTER T. & MEKKI. W. (2004). The penn arabic treebank : Building a large-scale annotated arabic corpus. *NEMLAR International Conference on Arabic Language Resources and Tools*.

MUTAWA A., ALNAJEM S. & ALZHOURI F. (2008). An hpsg approach to arabic nominal sentences. *JASIST*, **59**(3).

PARMENTIER Y., KALLMEYER L., LICHTE T., MAIER W. & DELLERT J. (2008). TuLiPA : A Syntax-Semantics Parsing Environment for Mildly Context-Sensitive Formalisms. In *9th International Workshop on Tree-Adjoining Grammar and Related Formalisms (TAG+9)*, p. 121–128, Tübingen, Germany.

PETITJEAN S. (2014). *Génération Modulaire de Grammaires Formelles*. PhD thesis, Université d'Orléans, France.

WEIR D. J. (1988). *Characterizing Mildly Context-Sensitive Grammar Formalisms*. PhD thesis, Université de Pennsylvanie, États-Unis.

XTAG RESEARCH G. (2001). A lexicalized tree adjoining grammar for english. *Technical Report IRCS-01-03, IRCS, University of Pennsylvania*.

# Déterminants et quantificateurs généralisés dynamiques

Clément Beysson[1]
(1) Université de Lorraine, 34 Cours Léopold, 54000 Nancy, France
clement.beysson@inria.fr

RÉSUMÉ

Dans cet article, nous proposons une classification des déterminants en étudiant leur capacité à introduire de nouveaux référents du discours et l'accessibilité de ces référents. Cette classification se fonde sur des aspects de logique dynamique (Groenendijk & Stokhof, 1991) dans la tradition montagovienne. Nous montrons ensuite que ces classes raffinent d'autres classifications plus linguistiques en étudiant chaque espèce de déterminants une à une. L'analyse de ces propriété est un première étape dans la définition des quantificateurs généralisés dynamiques nécessaires pour dénoter la sémantique des déterminants.

ABSTRACT

**Determiners and dynamic generalised quantifiers**

In this article, we propose a classification determiner classification by studying their ability to introduce new discours referents and these referents accessiility. This classification is based on dynamic properties from dynamic predicate logic (Groenendijk & Stokhof, 1991) in a montagovian tradition. By studying determiners speces, we show that these classes rafine some other linguistic classes. the analysis of these properties a step to define generalised dynamic quantifiers in order to denote the determiners semantic.

MOTS-CLÉS : Déterminants, Quantificateurs généralisés, Sémantique, Logique dynamique.

KEYWORDS: Determiners, Generalized quantifiers, Semantic, Dynamic Logic.

## 1 Introduction

Cet article propose de définir une classification des déterminants afin de s'en servir pour formuler dans de futurs travaux une formalisation de la sémantique de ces déterminants. elle se fonde sur une analyse des propriétés dynamiques des déterminants.

Pour cela, nous commençons dans la partie 2 par présenter le cadre formel dans lequel nous nous plaçons. Il s'agit d'un cadre montagovien (Montague, 1973) dans lequel les déterminants sont dénotés par des quantificateurs généralisés (Barwise & Cooper, 1981). Afin de ne pas se limiter à la sémantique des phrases et de tenir compte de celle des discours, nous étendons ensuite ce cadre à l'aide de la logique dynamique (Groenendijk & Stokhof, 1991) tout en définissant les trois dynamicités à la base de notre classification : la dynamicité interne, la dynamicité externe et la dynamicité intrinsèque. Nous poursuivons par un interprétation linguistique des trois dynamicités, ce qui nous permet de disposer de tests à effectuer sur les déterminants.

Ces fondements formels posés, nous définissons ce qu'est un déterminant dans la partie 3 en nous

concentrant sur le cas du français. Nous nous basons sur des grammaires générales comprenant une approche sémantique (Grevisse & Goosse, 2001; Riegel *et al.*, 2009). Nous reprenons ensuite une à une les diverses espèces de déterminants. Chacune est définie puis confrontée à nos tests pour en évaluer ses propriétés dynamiques. Rassembler les résultats nous permet de discuter les correspondances et divergences entre les classes usuelles de déterminants et nos dynamicités. De plus, toutes les combinaisons de dynamicité ne sont pas présentes, ce qui nous permet de définir seulement trois classes dynamiques de déterminants : les quantificateurs, les spécifiques et les généraux.

Nous revenons dans notre conclusion (partie 4) sur les limites de cette approche et sur les futurs travaux qui découlent du présent article.

# 2 Quantificateurs généralisé, déterminants et discours

## 2.1 Quantificateurs généralisés

Dans la tradition montégovienne (Montague, 1973), la sémantique des énoncés s'obtient de manière compositionnelle à partir de la sémantique de ses constituants et de leur organisation syntaxique. Ainsi, la dénotation des phrases suivantes s'obtient en composant la dénotation des syntagmes nominaux avec celle de *être gris*.

(1) a. **Un chat** est gris
b. **Tout chat** est gris
c. **Plus de la moitié des chats** sont gris

Plusieurs articles (Montague, 1973; Barwise & Cooper, 1981) proposent d'interpréter un syntagme nominal par un ensemble de prédicats. Cet ensemble de prédicats est l'ensemble des propriétés caractérisant le syntagme [1]. Il s'agira pour (1a) de tous les ensembles contenant au moins un chat. On pourra alors savoir si (1a) est vrai en regardant si l'ensemble de tout ce qui est gris appartient à la dénotation de *un chat*. Nous reviendrons plus tard sur les autres exemples.

D'un point de vue formel, il s'agit de quantificateurs généralisés, qui sont par définition des ensembles d'ensembles . Avec $e$ le type des individus et $t$ celui des valeurs de vérité, un syntagme nominal est donc de type $(e \to t) \to t$. Le nom lui-même peut être interprété par un prédicat [2]. Le déterminant se combinant avec lui pour former un syntagme nominal, il est donc de type $(e \to t) \to (e \to t) \to t$. Ce type, correspondant aux quantificateurs généralisés binaires, est celui des relations entre ensembles. Il nous indique donc comment la dénotation du nom et celle du syntagme verbal peuvent être reliées [3]. Par exemple, en reprenant les déterminants de (1), on obtient les dénotations suivantes :

(2) a. $\|un\|NV \equiv N \cap V \neq \emptyset$
*un chat est gris* ssi il existe un individu qui est à la fois un chat et gris.

1. Chaque propriété peut être modélisée comme un ensemble ou comme un prédicat, ce dernier étant alors la fonction caractéristique de l'ensemble.

2. Cette dénotation n'est en revanche valable que pour le cas des noms comptables. En effet, il n'est pas possible d'interpréter un nom massique comme un prédicat puisqu'un tel nom ne correspond pas à des individus. Les noms massiques seront donc exclus de la suite de cet article car leur interprétation est hors du cadre sur lequel nous nous basons.

3. Les syntagmes verbaux ne s'interprètent comme des prédicats que s'il s'appliquent à des individus. Ainsi, les verbes s'interprétant de manière collective et non distributive (Champollion, 2014) sorte du cadre de ces travaux.

b. $\|$*tout*$\|NV \equiv N \subset V$
*tout chat est gris* ssi tous les individus qui sont des chats sont également des individus gris.
c. $\|$*plus de la moitié des*$\|NV \equiv |N| < 2|N \cap V|$
*plus de la moitié des chats sont gris* ssi plus de la moitié des chats sont des chats gris.

Les quantificateurs généralisés étant très nombreux, ils sont capables d'exprimer des sémantiques très riches dont celle de (2c) qui exprime un déterminant qui n'est pas toujours considéré comme un déterminant syntaxique, ni même comme un constituant du discours. Les déterminants ne sont cependant pas les seuls à avoir des quantificateurs généralisés comme dénotation. Par exemple, le terme *les cartes des* peut également être interprété par un quantificateur généralisé dans l'énoncé suivant :

(3) **Les cartes des** joueurs sont posées sur la table

Afin de distinguer les quantificateurs généralisés dénotant des déterminants de ceux qui dénotent autre chose, on utilise la conservativité :

(4) Un déterminant $\mathcal{D}$ est conservatif ssi $\forall AB.\mathcal{D}AB \Leftrightarrow \mathcal{D}A(A \cap B)$

L'hypothèse a été émise (Barwise & Cooper, 1981) puis démontrée (Keenan & Stavi, 1986) que tous les déterminants sont conservatifs et qu'il s'agit d'une caractérisation des déterminants. Cela nous permet de rejeter *les cartes des* comme étant un déterminant car il n'est pas conservatif.

## 2.2 Quantificateurs généralisés et logique dynamique

Les phrases sont rarement isolées et fournir la dénotation d'un énoncé par une valeur de vérité est problématique. Par exemple, les références anaphoriques rendent impossible l'analyse indépendante des énoncés contenant des pronoms dont l'antécédent se trouve dans d'autres phrases. Si les deux phrases en (5) se suivent, alors la sémantique de (5b) n'est pas indépendante ; Elle dépend de celle de (5a) pour résoudre son anaphore.

(5) a. Il y a un chat dans la rue
b. Il miaule

Afin de gérer correctement cela, la Théorie de Représentation du Discours (DRT) fut développée (Kamp & Reyle, 1993). Ce formalisme propose un représentation récursive (DRS) de la sémantique d'un discours sous forme d'un couple. D'un côté les variables accessibles (correspondant aux référents introduits dans le discours) et de l'autre des formules logiques construites sur des variables ou sur des sous-DRS.

Pour fournir un fondement logique à la DRT, une logique dynamique a été développée (Groenendijk & Stokhof, 1991). Dans ce cadre, un énoncé est interprété comme une relation entre des contextes d'entrée et des contextes de sortie. Cela permet de modéliser l'évolution des références et des connaissances. Les contextes étant dans ce cadre des fonctions d'évaluation des variables libres, il est possible d'exprimer les effets dynamiques. Par exemple, la phrase (5b) modifiera ses contextes d'entrée pour garantir que la variable correspondant au *il* soit évaluée de la même manière que celle exprimant le *chat* dans (5a).

Pour capturer la manière dont les opérateurs logiques (conjonction, disjonction, implication …)

interagissent avec les propriétés dynamiques de leurs arguments (leur capacité à capturer des variables éloignées), deux notions furent introduites (Groenendijk & Stokhof, 1991). Un connecteur binaire est défini comme dynamique interne si les lieurs présents dans son premier argument peuvent lier des variables apparaissant dans son second argument. De même, un opérateur binaire est défini comme dynamique externe si les lieurs présents dans ses arguments peuvent lier des variables présentes en dehors de l'opérateur. On peut par exemple voir formellement que les propriétés de la conjonction (6a,6b) et de la disjonction (6c,6d) en logique dynamique ont pour conséquences que le premier est dynamique interne et externe quand le second n'est aucun des deux.

(6) a. $(\exists x.\phi) \wedge \psi \equiv \exists x.(\phi \wedge \psi)$
b. $((\exists x.\phi) \wedge \psi) \wedge \xi \equiv \exists x.(\phi \wedge \psi \wedge \xi)$
c. $(\exists x.\phi) \vee \psi \not\equiv \exists x.(\phi \vee \psi)$
d. $((\exists x.\phi) \vee \psi) \vee \xi \equiv \exists x.(\phi \vee \psi \vee \xi)$

Les quantificateurs généralisés binaires pouvant s'exprimer en logique dynamique, on peut s'interroger sur leur dynamicité. C'est surtout leur dynamicité interne qui a été étudiée (Kanazawa, 1994; Chierchia, 1992) et nous y reviendrons en (2.3). Cependant, contrairement aux connecteurs logiques, les quantificateurs généralisés dynamiques peuvent également lier des variables eux-même. Cela introduit une troisième dynamicité : la dynamicité intrinsèque. Un quantificateur généralisé dynamique est dynamique intrinsèque si il peut lier des variables en dehors de ses arguments.

## 2.3 Propriétés dynamiques des déterminants

Comme nous l'avons vu, les quantificateurs généralisés dynamiques peuvent avoir trois types de dynamicité (interne, externe, intrinsèque). Afin de pouvoir, dans de futurs travaux, proposer des quantificateurs dénotant des déterminants, nous allons à présent étudier la présence ou l'absence de ces trois dynamicités pour les déterminants. Il nous faut donc établir des tests linguistiques pour les déterminants reflétant les propriétés de liaisons de variables à la base de la dynamicité.

D'un côté, le déterminant met en relation le sens du syntagme nominal (appelé restricteur) et celui du syntagme verbal (appelé portée). De l'autre, la dynamicité interne correspond à la possibilité pour le premier argument du connecteur de lier les variables du second. Cette dynamicité s'interprète donc linguistiquement comme la possibilité de faire référence dans la portée à une entité introduite au sein du restricteur. Nous proposons de vérifier cette possibilité comme suit :

(7) a. **Un** joueur qui pioche [un As]$_i$ [le]$_i$ retourne sur la table
b. **tout** joueur qui pioche [un As]$_i$ [le]$_i$ retourne sur la table
c. **Plus de la moitié des** joueurs qui piochent [un As]$_i$ [le]$_i$ retournent sur la table

Ces trois énoncés étant corrects, on peut en déduire que *un*, *tout* et *plus de la moitié des* sont dynamique interne.

Pour la dynamicité externe, les arguments peuvent lier des variables à l'extérieur du connecteur. Cela s'interprète linguistiquement par la possibilité de faire référence dans la suite du discours à une entité introduire dans le restricteur ou dans la portée. Exprimons cette propriété par une phrase :

(8) a. Si **Un** joueur pioche [un As]$_i$, le croupier [le]$_i$ retirera du jeu
b. *Si **tout** joueur pioche [un As]$_i$, le croupier [le]$_i$ retirera du jeu
c. Si **Plus de la moitié des** joueurs piochent [un As]$_i$, le croupier [le]$_i$ retirera du jeu

Cette fois, l'énoncé (8b) n'est pas correct alors que les deux autres le sont. Cela nous montre que *tout* n'est pas dynamique externe, contrairement aux deux autres.

Un connecteur est dynamique intrinsèque s'il peut lui-même lier des variables en dehors de ses arguments. Linguistiquement, cela correspond à la possibilité de faire référence à l'entité actualisée par le déterminant dans un autre énoncé. Les énoncés suivants correspondent à cette propriété :

(9) a. Si [**Un** joueur]$_i$ pioche un As, [il]$_i$ remporte le point
b. *Si [**tout** joueur]$_i$ pioche un As, [il]$_i$ remporte le point
c. Si [**Plus de la moitié des** joueurs]$_i$d piochent un As, [ils]$_i$ remportent le point

De nouveau, *tout* échoue à notre test et rend (9b) incorrect. Nous en déduisons qu'il n'est pas intrinsèquement dynamique alors que les deux autres déterminants le sont.

De manière général, nous pouvons maintenant vérifier la dynamicité de n'importe quel déterminant $\mathcal{D}$ en évaluant la correction de trois énoncés [4] :

(10) a. (Interne) $\mathcal{D}$ joueur(s) qui pioche(nt) [un As]$_i$ [le]$_i$ retourne(nt) sur la table
b. (Externe) Si $\mathcal{D}$ joueur(s) pioche(nt) [un As]$_i$, le croupier [le]$_i$ retirera du jeu
c. (Intrinsèque) Si [$\mathcal{D}$ joueur(s)]$_i$ pioche(nt) un as, [il(s)]$_i$ remporte(nt) le point

A priori, nos trois tests devraient mener à $2^3 = 8$ classes de déterminants. Afin de vérifier si toutes ces classes sont habitées par des déterminants et si ces classes sont corrélées avec des classes grammaticales, nous allons avoir besoin de disposer d'une liste de déterminants et d'une définition des classes.

# 3 Les déterminants du Français

Les éléments formels présentés jusqu'ici ne sont pas spécifiques à une langue en particulier, mais nous avons choisi d'utiliser spécifiquement le français pour la suite afin de nous permettre de juger plus facilement la correction des énoncés lors des tests linguistiques. De futurs travaux permettront de confirmer si les résultats que nous obtenons sont généralisables à toutes les langues. Les définitions et classifications des déterminants étant très diverses, nous avons choisi de nous baser sur deux grammaires générales et reconnues du français qui insistent sur les aspects sémantiques. Il s'agira de (Grevisse & Goosse, 2001) et de (Riegel *et al.*, 2009).

## 3.1 Définition

Le groupe nominal est un type de syntagme central dans la structure des phrases. En effet, ce sont ces syntagmes qui désignent les entités dont on souhaite parler en les nommant et en les qualifiant. Le cœur de ce groupe est a priori le nom, ce dernier étant ensuite modulé par divers modifieurs du nom (adjectifs, groupes prépositionnels, propositions relatives) permettant d'en affiner le sens. Cependant, selon (Riegel *et al.*, 2009), le déterminant est le mot qui doit nécessairement précéder un nom commun pour constituer un groupe nominal bien formé. Nous disposons donc d'une caractérisation

4. Il faut noter que le test de dynamicité intrinsèque est nécessaire mais pas suffisant. En effet, il ne permet pas de distinguer cette dynamicité intrinsèque, où le déterminant produit une référence accessible, des cas où la référence est produite ailleurs et est accessible aussi bien pour le déterminant que pour la suite.

distributionnelle des déterminants. On observe alors que la classe des déterminants est partiellement ouverte. En effet, une grande partie de ses éléments forme une classe fermée bien définie tandis qu'une frange, appelée déterminants occasionnels par (Grevisse & Goosse, 2001), forme une classe ouverte se nourrissant d'autres catégories morpho-syntaxiques.

De plus, le nom et ses modifieurs ne suffisent pas à désigner une entité du discours et restent une description abstraite d'un concept. En effet, le mot *chat* ne désigne rien d'autre que la propriété d'être un chat, sans que cela ne corresponde au moindre référent. C'est donc le rôle sémantique du déterminant de faire passer le nom et ses modifieurs du monde des idées à celui des référents du discours. On parle alors d'actualisation. Un nombre restreint de cas particuliers semble ne pas utiliser de déterminant pour cette actualisation et on parle alors d'article absent. Un effet supplémentaire de cette actualisation est de pouvoir actualiser n'importe quelle entité linguistique, qui sera alors interprétée comme un nom. On peut ainsi tout aussi bien actualiser un adjectif (11a) qu'une lettre (11b), voir même qu'une phrase (11c).

(11) a. *Le* **blanc** te va bien
b. *Le* **e** est absent de ce poème
c. Les électeurs se lassent *des* **je vous le promets**

On pourrait argumenter pour (11a) et (11b) qu'il s'agit déjà de nom, et cela est vrai d'un point de vu distributionnel, mais (11c) montre bien qu'un objet qui n'est clairement pas un nom peut être actualisé comme s'il en était un. Il nous semble donc pertinent d'analyser de la même manière (11a) et (11b) et de préciser qu'ils sont interprétés comme des lettres/adjectifs et comme des noms car ils sont suffisamment souvent actualisés comme des noms pour être interprétés comme tel. Cela complique le typage des déterminants dans le sens où ils considèrent tout objet linguistique comme un nom.

## 3.2 Espèces de déterminant

Intéressons-nous aux diverses espèces de déterminants [5] (Grevisse & Goosse, 2001), séparées par des distinctions de distribution et de sémantique. Chacune de ces espèces recouvre plusieurs mots-formes mais nous ne nous étendrons pas sur ces variations morphologiques et nous nous contenterons d'exemples. Nous nous concentrerons en revanche sur la sémantique de ces déterminants, dans le sens où l'actualisation des noms n'est pas neutre et chaque déterminant intègre différemment le nom qu'il détermine dans l'énoncé.

### 3.2.1 L'article défini

Voici quelques formes de l'article défini :

(12) a. Fais attention à *la* route
b. *L*'homme est mortel
c. Je n'ai pas encore lu *les* livres que tu m'as offerts

Afin que l'article défini puisse actualiser un nom, il est nécessaire que ce nom puisse correspondre à une entité présente dans le contexte d'énonciation et connu de tous les interlocuteurs. Le syntagme

5. Nous ignorerons volontairement les espèces des déterminants interrogatifs et exclamatifs, ces derniers sortant du champ de notre étude car n'apparaissant pas dans des énoncés déclaratifs. De même, l'article absent ne sera pas abordé car il ne s'agit a priori pas d'un déterminant mais bien de son absence.

nominal fait alors référence à cette entité. Il est de plus présupposé qu'il y a une unicité de cette entité dans le contexte, ou du moins qu'une occurrence de cette entité est saillante dans le contexte par rapport aux autres à l'instant de l'énonciation. Dans le cas du pluriel, c'est un sous-ensemble d'entité qui est saillant. Deux lectures sont alors possibles. Soit il est fait référence à une entité précise correspondant au nom (lecture spécifique)(12a), soit il est fait référence à l'ensemble de toutes les entités correspondant au nom (lecture générique)(12b).

Les tests montrent que l'article défini est dynamique interne et externe. Concernant la dynamicité intrinsèque, l'article défini valide notre test mais il s'agit d'un faux positif comme expliqué précédemment. En effet, il ne produit pas lui-même de référence et la suite du discours a accès à la même référence antérieure que lui. Il n'est donc pas intrinsèquement dynamique.

### 3.2.2 L'article indéfini et partitif

Commençons par préciser que l'article partitif s'applique par définition à des noms massiques. Ces derniers étant exclus de notre analyse, nous ne parlerons pas non plus de ces déterminants. Plusieurs formes de l'article indéfini sont possibles :

(13) a. *Un* chat retombe toujours sur ses pattes
b. Je pars acheter *des* livres de Jules Verne
c. J'aimerais bien trouver *une* licorne

L'article indéfini actualise les noms en produisant de nouvelles entités correspondant à ce nom sans pour autant l'identifier. Il s'agira alors d'un nouveau référent utilisable dans la suite du discours. Plusieurs lectures de cet article sont alors possibles, changeant l'interprétation de cette nouvelle entité.

Tout d'abord, il peut s'agir d'une lecture dite spécifique, dans le sens où même si l'entité n'est pas identifiée, elle pourrait l'être car elle existe bel et bien (13b). Dans ce cas, l'article indéfini se contente d'affirmer l'existence d'une occurrence du nom. Il est néanmoins possible que cette existence ne soit que virtuelle et n'implique pas l'existence de quoi que ce soit. L'article se contente alors de l'introduire comme pouvant exister et permet d'y faire référence par la suite (13c). Une lecture générique est également possible. Dans ce cas, c'est un représentant quelconque du nom qui est introduit afin de parler par son intermédiaire de tout ce que recouvre le nom. Cette lecture n'est possible qu'au singulier.

Lorsque l'on vérifie la classe dynamique de l'article indéfini, on trouve qu'il dispose de toutes les dynamicités.

### 3.2.3 Le déterminant numéral

Le déterminant numéral contient des formes bien plus riches que les articles :

(14) a. Cela coûte *deux cent quatre-vingts-trois* euros
b. *Un* tiens vaut mieux que *deux* tu l'auras
c. Je compte acheter *entre quatre et six* pommes
d. J'ai vu *environ trente* chats hier

Les déterminants numéraux ont une sémantique proche de l'article indéfini mais ajoutent une notion de cardinalité, cette dernière n'étant pas nécessairement exacte (14c,14d). On retrouve en effet les

diverses lectures (spécifique (14a) ou générique (14b)) ainsi que l'introduction de nouvelles entités dans le contexte du discours. Cependant, le déterminant numéral n'introduit pas d'entités mais des ensembles d'entités dont le cardinal est renseigné.

En appliquant les tests de dynamicité à ces déterminants, on se rend compte qu'ils sont dynamiques internes, externes et intrinsèques, exactement comme l'article indéfini.

### 3.2.4 Le déterminant possessif

La morphologie du déterminant possessif n'est pas seulement fonction du nom qu'il détermine mais également d'un antécédent jouant le rôle sémantique du possesseur.

(15) a. Je te prête *mon* livre
b. *Mes* sœurs sont trop petites pour m'accompagner
c. *Je te prête le livre de moi
d. **Leurs* chiens des voisins aboient souvent
e. Il a perdu *la* mémoire

Cependant, cette notion de possesseur n'est pas une relation de propriété. Elle marque plutôt une relation, variable en fonction du contexte. (15a) pourrait être *le livre que j'ai écrit* tout comme *le livre que j'ai acheté*. De même, (15b) n'indique pas d'appartenance. Plus généralement, la sémantique de (15a) et de (15c) est identique, bien que la seconde ne soit pas acceptable (le même procédé est applicable aux autres personnes). Ainsi, il n'est pas possible de faire le cumul avec un complément qui identifierait le possesseur comme dans (15d) et il est courant (voire obligatoire) de remplacer le déterminant possessif par un simple article défini lorsque le possesseur est évident (15e).

S'agissant d'une variation de l'article défini, le déterminant possessif est lui aussi seulement dynamique interne et externe.

### 3.2.5 Le déterminant démonstratif

Le déterminant démonstratif permet de faire référence à une entité que l'on choisit de rendre saillante dans le contexte. Cette saillance peut être accompagnée d'un geste s'il s'agit de faire ressortir l'entité dans l'environnement spatial des locuteurs.

(16) a. *Cet* homme est dangereux
b. *Ces* voitures-ci doivent coûter cher
c. Je veux *cette* pomme-là

Le déterminant démonstratif permet de faire référence à trois catégories d'entités. Tout d'abord, il peut s'agir d'une anaphore désignant une entité déjà exprimée dans le discours. Plus rarement, il est également possible de l'employer de manière cataphorique afin de désigner une référence à venir dans le discours. Il est enfin possible de s'en servir pour désigner une entité qui n'est pas présente dans le discours mais dans le contexte d'énonciation. Bien que les locuteurs eux-mêmes soient clairement saillants dans tout contexte d'énonciation, le déterminant démonstratif ne fait pas référence à ces derniers et donne une notion de tiers. Par exemple, (16a) ne fera pas référence à l'un des interlocuteurs, sauf de manière péjorative ou pour clôturer la conversation en indiquant que la personne désignée n'est justement plus un interlocuteur.

Contrairement à l'article défini, qui fait également référence à une entité saillante du contexte, le déterminant ne présuppose pas l'unicité de cette entité et c'est précisément le rôle de ce déterminant que de mettre en avant au sein du contexte une des entités décrites par le nom. Dans le cas où il est fait référence à une entité qui est effectivement unique, le déterminant démonstratif et l'article défini ont la même sémantique.

Il découle des tests que le déterminant démonstratif est, comme l'article défini, dynamique interne et externe, mais pas dynamique intrinsèque.

### 3.2.6 Le déterminant relatif

Le déterminant relatif est une forme archaïque qui n'existe presque plus à l'oral et ne subsiste à l'écrit que dans des domaines précis comme les textes juridiques.

(17) a. Il a accusé la fille de la prévenue, *laquelle* fille était enceinte
b. Il a accusé la fille de la prévenue, et cette fille était enceinte

Le déterminant relatif permet de faire une référence explicite à un antécédent construit sur le même nom que celui que complète le déterminant. Il s'agit encore une fois de ne pas produire de nouvelles entités. La différence étant que cette fois il n'est pas possible de faire appel à autre chose qu'une entité introduite précédemment dans le discours. Ce déterminant est donc sémantiquement équivalent à l'utilisation de la conjonction *et* suivit du déterminant démonstratif comme dans (17b) et est donc un quantificateur très similaire. A l'écrit, cette forme n'est principalement utilisée que lorsque l'utilisation du pronom relatif correspondant conserverait une ambiguïté qu'il est possible de lever avec le déterminant relatif comme dans (17a).

Les tests montrent que ce déterminant est dynamique interne et externe mais pas intrinsèque.

### 3.2.7 Le déterminant indéfini

Il s'agit ici de l'espèce de déterminant la plus variée. En effet, une partie d'entre eux forme une classe fermée tandis que l'autre dispose de règles de construction en faisant une classe ouverte appelée déterminants indéfinis occasionnels (les mots le composant n'étant usuellement pas des déterminants)(Grevisse & Goosse, 2001). Il n'est donc pas possible de tous les lister et d'en décrire le sens précis. On peut tout de même commencer par lister les déterminants indéfinis qui ne sont pas occasionnels. Il s'agit des lemmes suivants :

*aucun*, *certain*, *chaque*, *différents*, *divers*, *maint*, *nul*, *plusieurs*, *quelques*, *tel*, *tout*.

A cela, on peut ajouter un grand nombre d'autres expressions polylexicales que l'on considérera comme des déterminants occasionnels.

(18) a. J'ai acheté *très peu de* pommes aujourd'hui
b. *Un tas de* français se sont abstenus
c. Je lui ai déjà dit *je ne sais combien de* fois de ne pas y aller
d. J'ai lu ce livre *x* fois, j'en ai assez

On trouve parmi ces déterminants indéfinis occasionnels des formes aussi variées que l'utilisation d'adverbes de quantité ou de degré (18a), de noms donnant une notion de quantité (18b), de phrases

complètes (18c) mais aussi de termes mathématiques passés dans le langage courant (18d).

Une grande difficulté que l'on rencontre avec ces déterminants occasionnels est qu'il n'est pas toujours évident de les identifier comme tel. En effet, il ne s'agit presque jamais de mots mais d'expressions polylexicales que tout le monde ne considère pas comme des constituants. Comme il ne s'agit pas d'une classe fermée, il n'est pas possible d'en faire la liste. Ils sont de plus construits sur des mots appartenant à d'autres classes, on pourrait argumenter qu'il ne s'agit aucunement de déterminants mais bien d'adverbes, de noms, de phrases ...

Il est cependant possible de voir que d'un point de vue distributionnel, il s'agit bien de déterminants. Ensuite, nous nous intéressons plus particulièrement à la sémantique des déterminants et de ce point de vue, ces expressions actualisent bien le nom qu'elles déterminent. On remarquera également que (18b) présente un accord du verbe avec le nom qui est actualisé et non avec le nom inclus dans le déterminant, qui devrait pourtant être le sujet syntaxique.

(19) a. *Un* tas de sable s'est formé
b. *Les* familles des patients ne sont pas admises

Toujours en se basant sur (18b), on peut voir que nos critères de définition pour les déterminants sont sémantiques et non syntaxiques, puisque (19a) et (19b) disposent d'une structure syntaxique apparemment identique mais sémantiquement différente. Pour (19a), la notion de *tas* est littérale et non imagée (il s'agit effectivement d'un tas), ce qui pousse à ne pas considérer *un tas de* comme un déterminant. Bien plus flagrant, le cas de (19b) montre clairement qu'on ne peut pas trop généraliser la construction des déterminants indéfinis occasionnels afin d'éviter de considérer que *les familles des* est un déterminant.

D'un point de vue dynamique, les tests montrent qu'ils sont tous dynamiques internes, mais qu'il y a une certaine variabilité pour les autres dynamicités. Par exemple, *certains* est à la fois dynamique externe et intrinsèque alors que *tout* n'est que dynamique interne [6].

### 3.2.8 La composition des déterminants

Au-delà de ces espèces de déterminants, on observe également qu'il est possible de composer plusieurs déterminants d'espèces potentiellement différentes afin de produire un nouveau déterminant. On peut entre autres faire suivre *tout* d'un article défini, d'un déterminant possessif ou d'un déterminant démonstratif (*tous les*, *toute cette*, ...). Ces mêmes articles définis, déterminants possessifs et déterminants démonstratifs peuvent également être suivis d'un déterminant numéral ou d'un déterminant indéfini (*les trois*, *ces divers* ...) On observe aussi qu'il est possible d'ajouter *même* ou *autre* à un déterminant pour former un déterminant composé ajoutant la notion d'identité ou de disparité.

Dans tous ces cas, il est possible d'interpréter la composition de déterminants comme un seul déterminant dont le sens est la composition des sens de ses composants. l'étude de la dynamicité de ces composés fera l'objet de futur travaux.

6. Bien que la classe des déterminants indéfinis soit ouverte et qu'une exception ai pu nous échapper, nous conjecturons qu'il n'y a pas de déterminant indéfini qui soit dynamique externe mais pas intrinsèque ou dynamique intrinsèque mais pas dynamique.

## 3.3 Classification des déterminants

Les diverses espèces de déterminants sont regroupées en deux grandes classes (Riegel *et al.*, 2009) : les déterminants définis et les déterminants indéfinis.

La classe des déterminants définis regroupe l'article défini, le déterminant possessif et le déterminant démonstratif. L'article défini et le déterminant démonstratif sont capables de faire une référence anaphorique, ce qui les rend définis. Comme indiqué en (3.2.4), le déterminant possessif est sémantiquement la composition de l'article défini avec le pronom se référant au possesseur et fait donc partie de la même classe que l'article défini. La classe des indéfinis est alors identifiée par défaut comme l'ensemble des espèces ne permettant pas la référence anaphorique. Cela cause un problème puisque l'espèce des déterminants relatifs est par nature une référence anaphorique et qui est sémantiquement proche du déterminant démonstratif (3.2.6) mais n'est pas comptée parmi les définis.

Une autre classification (Grevisse & Goosse, 2001) propose de regrouper l'article indéfini, les déterminants numéraux et certains déterminants indéfinis dans la classe des quantificateurs tandis que les autres forment la classe des identifiants. Sémantiquement, il s'agit de savoir si l'actualisation du nom par le déterminant correspond à une entité identifiée du contexte ou non.

Comme les classifications usuelles portent sur des propriétés sémantiques de l'actualisation et des références et qu'il s'agit justement d'éléments exploités par notre approche formelle des déterminants, nous pouvons les comparer à celle résultant de notre formalisme dans (TABLE 1).

| | + − *défini indéfini* | + − *quantificateur identifiant* | *dynamique interne* | *dynamique externe* | *dynamique intrinsèque* | *exemple* |
|---|---|---|---|---|---|---|
| *article défini* | + | − | + | + | − | le |
| *article indéfini* | − | + | + | + | + | un |
| *déterminant numéral* | − | + | + | + | + | trois |
| *déterminant possessif* | + | − | + | + | − | son |
| *déterminant démonstratif* | + | − | + | + | − | ce |
| *déterminant relatif* | − | − | + | + | − | lequel |
| *déterminant indéfini* | − | ± | + | + | + | certain |
| | | | + | − | − | aucun |

TABLE 1 – Comparatif de diverses classes de déterminants

Remarquons que la classe des définis et celle des identifiants sont presque identiques, tout comme celles des indéfinis et des quantificateurs. Ils se distinguent par le déterminant relatif, qui devrait selon nous être défini, et pas un déterminant indéfini (dont l'hétérogénéité explique qu'il ne soit pas simple à classer).

Du point de vue dynamique, tous les déterminants sont dynamiques internes. Il s'agit donc d'une propriété qu'ils partagent tous et qui peut nous aider à les caractériser formellement. Il y a de plus une équivalence entre défini, identifiant et non dynamique intrinsèque (les déterminants indéfini

mis à part). La dynamicité externe n'a quant à elle pas d'effet sur la classification classique des déterminants.

Comme le montre (TABLE 1), il n'existe donc que trois classes dynamiques de déterminants au lieu des huit envisagées. Nous proposons donc de les classer en déterminants quantificateurs, spécifiques et généraux comme définit dans (TABLE 2).

| *déterminants* | *dynamique interne* | *dynamique externe* | *intrinsèquement dynamique* | *exemple de déterminant* |
|---|---|---|---|---|
| *quantificateurs* | + | − | − | *tout* |
| *spécifiques* | + | + | − | *ce* |
| *généraux* | + | + | + | *un* |

TABLE 2 – Classes dynamiques de déterminants

# 4 Conclusion

Grâce à nos trois tests linguistiques de dynamicité, nous connaissons désormais les propriétés dynamiques des déterminants. En particulier, nous avons montré qu'il n'y avait pas huit mais seulement trois classes de déterminants selon nos critères. Bien que notre étude se soit concentrée sur le français, nous conjecturons que ces résultats se généralisent à d'autres langues.

Pour produire des dénotations sémantiques pour les déterminants, nous savons désormais qu'il suffit de produire des quantificateurs généralisés dynamiques appartenant à l'une de nos trois classes. Si on ajoute la conservativité (Keenan & Stavi, 1986), nous obtenons une caractérisation assez précise des quantificateurs permettant d'exprimer la sémantique des déterminants.

De nombreuses limitations restent tout de même présentes. Nous ne traitons en effet pas les déterminants s'appliquant à des noms massiques car notre modèle se base sur la notion d'individu, absente dans les objets non comptables. De plus, notre gestion du pluriel est limitée car nous ne prenons en compte que des cas distributifs (la portée se distribue sur toutes les entités composant la portée) et non collectifs ou cumulatifs (Champollion, 2014). En effet, il serait pour cela nécessaire de passer à l'ordre supérieur afin d'appliquer le prédicat dénotant le syntagme verbal à un ensemble d'individu et non pas à un unique individu.

# Références


BARWISE J. & COOPER R. (1981). Generalized quantifiers and natural language. *Linguistic & Philosophy*, **4**(2), 159–219.

CHAMPOLLION L. (2014). Distributivity, collectivity and cumulativity. *Wiley's companion to semantics.*

CHIERCHIA G. (1992). Anaphora and dynamic binding. *Linguistic & Philosophy*, **15**(2), 111–183.

GREVISSE M. & GOOSSE A. (2001). *Le Bon Usage*. Duculot, 13e edition.

GROENENDIJK J. & STOKHOF M. (1991). Dynamic predicate logic. *Linguistic & Philosophy*, **14**(1), 39–100.

KAMP H. & REYLE U. (1993). *From Discourse to Logic*. Dordrecht : Kluwer Academic Publishers.

KANAZAWA M. (1994). Dynamic generalized quantifiers and monotonicity. In M. KANAZAWA & C. PIÑÒN, Eds., *Dynamics, Polarity and Quantification*, volume 48 of *CSLI Lecture Notes*, p. 213–250. Stanford : CSLI Publications.

KEENAN E. L. & STAVI J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, **9**(3), 253–326.

MONTAGUE R. (1973). The proper treatment of quantification in ordinary English. In J. H. E.A., Ed., *Approaches to Natural Language*, p. 221–242. Reidel.

RIEGEL M., PELLAT J.-C. & RIOUL R. (2009). *Grammaire méthodique du français*. Linguistique nouvelle. Paris : Presses universitaires de France, 4e edition.

# Détection de l'incertitude et de la négation : un état de l'art

Clément Dalloux[1]
(1) IRISA - CNRS, Campus de Beaulieu, 35042 Rennes, France
clement.dalloux@irisa.fr

RÉSUMÉ

L'un des objectifs de nos travaux, à terme, est de transformer un corpus de documents médicaux en données structurées pour en faciliter l'exploitation. Ainsi, il est nécessaire non seulement de détecter les concepts médicaux évoqués, mais aussi d'intégrer un processus capable d'identifier le contexte dans lequel est évoqué chaque concept médical. Dans cet article, nous revenons principalement sur les systèmes par apprentissage supervisé qui ont été proposé pour la détection de l'incertitude et de la négation. Ces dix dernières années, les travaux pour détecter l'incertitude et la négation dans les textes en anglais ont donné des résultats satisfaisants. Cependant, il existe encore une marge de progression non-négligeable.

ABSTRACT

**Identifying uncertainty and negation's cues and scope : State of the art**

One of the goals of our endeavours is to turn a corpus of medical documents into more easily readable structured data. Thus, it is necessary to add a process capable of identifying the context in which each medical concept is used. In this article, we mainly review the various systems proposed for negation and uncertainty detection that include a machine learning method. This last decade, studies undertaken in order to detect uncertainty and negation in texts have given satisfying results, however, there is still room for improvement.

MOTS-CLÉS : étiquetage automatique, incertitude, négation, apprentissage artificiel, apprentissage supervisé.

KEYWORDS: automatic labeling, uncertainty, speculation, negation, machine learning, supervised learning.

## 1 Introduction

La détection d'évènements incertains et niés dans les textes est une thématique de recherche à laquelle les chercheurs en TAL s'intéressent depuis une quinzaine d'années. Cela est notamment le cas dans le domaine biomédical dans lequel s'inscrit nos travaux de thèse ; en effet, beaucoup d'informations sont apportées sous forme non structurée (c'est-à-dire sous la forme de texte libre) et il est important de savoir détecter les incertitudes et les négations qui sont utilisées pour exprimer des impressions, faire des hypothèses ou bien donner des résultats négatifs. D'ailleurs, l'élaboration du corpus BioScope ainsi que la tâche commune qui eut lieu lors de la *Conference on Computational Natural Language Learning* de 2010 (Farkas *et al.*, 2010) montrent l'intérêt de cette tâche pour la fouille de textes biomédicaux (cf. infra).

Cet article présente un état de l'art des travaux sur la détection automatique de la négation et de

l'incertitude, qui constitueront le point de départ de nos travaux. Ainsi, nous présentons les approches proposées ces dernières années selon deux axes, la détection des marqueurs, c'est-à-dire des termes qui expriment une négation ou une incertitude, et la détection de la portée, c'est-à-dire des termes qui sont impactés par ces phénomènes.

L'article est structuré de la manière suivante. Dans la section qui suit, nous définissons ces phénomènes, nous donnons une liste de marqueurs typiques, et nous analysons les problèmes liés à leur détection. Dans la section 3, nous présentons les ensembles de données qui ont été annotés dans le but d'automatiser leur détection, soit pour procéder à un entraînement (apprentissage supervisé) ou bien pour extraire divers types de descripteurs. Les techniques par système expert et apprentissage artificiel dédiés à la détection des marqueurs et de la portée sont abordées dans la section 4. Enfin, nous présentons nos conclusions et perspectives dans la dernière section.

## 2 Détection des phénomènes : Analyse du problème

Le Dictionnaire de l'Académie française (9ème édition) définit la négation comme *l'énoncé qui rejette comme faux, qui déclare faux, une proposition ou un jugement, ou bien, du point de vue de la linguistique, comme la construction utilisée pour nier ou pour exprimer le refus* et l'incertitude comme *le caractère de ce qui est incertain, douteux ou imprévisible*. Ces définitions génériques n'indiquent aucune piste pour permettre de repérer ces phénomènes dans les textes, mais laissent supposer une grande variété de réalisations linguistiques. En effet, ces constructions s'expriment par l'utilisation de marqueurs que l'on retrouve dans la plupart des catégories grammaticales. Nous présentons ci-dessous un ensemble de marqueurs de négation parmi les plus courants pour l'anglais, affixes compris, et des principaux marqueurs d'incertitude. Il s'agit bien sûr d'un ensemble non exhaustif et ouvert :

- **Négation**
  - Auxiliaire : *Cannot*.
  - Noms : *Absence, lack*.
  - Verbes : *lack, exclude, fail, favor over, miss, rule out*.
  - Conjonctions : *neither...nor, rather than*.
  - Adjectifs : *absent, impossible, negative, unable*.
  - Adverbes : *no, never, no longer, not*.
  - Prépositions : *except, instead of, with the exception of, without*.
  - Affixes : *un-, in-, im-, dis-, non-, -less, etc.*
- **Incertitude**
  - Auxiliaires : *may, might, can, would, could, should*.
  - Verbes : *suggest, question, presume, suspect, indicate, suppose, seem, appear, favor*.
  - Conjonctions : *or, and/or, either...or*.
  - Adjectifs ou adverbes : *problable, likely, possible, unsure*.

La principale difficulté pour la tâche de détection des marqueurs est d'être capable de résoudre les ambiguïtés, c'est-à-dire qu'il faut que le système soit capable de différencier les contextes où les termes sont utilisés comme marqueur de ceux où il ne le sont pas. Par exemple, l'auxiliaire *can* peut être utilisé sans marquer l'incertitude (1). Cela vaut pour la plupart des termes qui servent à exprimer l'incertitude.

1. It can only be activated in a tissue-specific manner by elements that lie downstream of the initiation site.

Dans la plupart des cas, l'effet du marqueur sur la phrase s'étend sur tout ou partie des unités lexicales. Celles impactées forment ce que l'on appelle communément **la portée** (*scope* en anglais). Identifier

automatiquement ces unités est une tâche complexe pour plusieurs raisons. Par exemple, la présence d'un ou plusieurs marqueurs de négation n'implique pas toujours que la phrase soit négative, elle peut être incertaine (1), voire positive si accompagnée par un adverbe de fréquence (2). La portée peut s'étendre des deux côtés du marqueur (3 et 4) et être discontinue (5 et 6). La méthode automatique devra donc être capable de capturer la dépendance entre le sujet et le prédicat nié, tout en excluant le syntagme verbal positif. Par ailleurs, plusieurs portées peuvent se chevaucher dans les énoncés qui contiennent plusieurs marqueurs (7 et 8). Ces portées doivent être repérées indépendamment l'une de l'autre.

1. *A small amount of adenopathy* ***cannot be*** *completely* ***excluded****.*
2. *Treated early, cancer is* ***not*** *always a death sentence.*
3. *Horseshoe* ***or*** *pelvic kidneys, ureteral abnormalities.*
4. *asthma* ***may*** *limit ability to participate in the exercise intervention.*
5. *The patient visited the hospital but did* ***not*** *meet the doctor.*
6. *The patient is coming to the hospital but* ***might*** *be late.*
7. *The patient did* ***not*** *take his medication because he was not feeling sick.*
8. *The patient did not take his medication because he was* ***not*** *feeling sick.*

Les publications liées à ces tâches de détection ou de classification ont principalement été publiées dans le cadre de la description du corpus BioScope (Vincze *et al.*, 2008), ainsi que lors des tâches communes de CoNLL-2010 (Farkas *et al.*, 2010) portant sur la détection de l'incertitude, et SEM-2012 (Morante & Blanco, 2012) portant sur la détection de la portée et de la cible de la négation. Nous présentons BioScope ainsi que les ensembles de données utilisés dans la section suivante.

# 3 Les données

Ces dernières années, avec la démocratisation des techniques d'apprentissage supervisé, de nombreux corpus de spécialité ont été annotés afin d'entraîner des modèles pour la détection automatique d'évènements et de phénomènes linguistiques. Cette section est dédiée à la description des données annotées pour l'incertitude et/ou la négation.

## 3.1 BioScope

Le corpus BioScope (Vincze *et al.*, 2008), est constitué de 3 sources textuelles différentes, à savoir de comptes-rendus d'examens radiologiques, d'articles de FlyBase et BMC Bioinformatics, ainsi que de résumés d'articles scientifiques issus du corpus GENIA [1]. Selon les consignes d'annotations [2] préétablies, cet ensemble de données est annoté au niveau des marqueurs de l'incertitude et de la négation, et au niveau de la phrase pour marquer leur portée linguistique. Les portées discontinues ne sont pas prises en compte. Le tableau 1 quantifie la composition du corpus, avec $Pneg$ phrases négatives, $M_{neg}$ marqueurs de la négation, $P_{inc}$ phrases incertaines, $M_{inc}$ marqueurs de l'incertitude.

Illustrées en figure 1, les données sont disponibles au format XML, où chaque phrase et chaque paire portée/marqueur (*xcope* et *cue*) sont indexées par un identifiant unique.

1. `http://www.geniaproject.org/`
2. `http://bit.ly/2mDnAoL`

| | Compte-rendus | Articles | Résumés |
|---|---|---|---|
| Documents | 1 954 | 9 | 1 273 |
| Phrases | 6 383 | 2 670 | 11 871 |
| $P_{neg}$ | 13,55% | 12,70% | 13,45% |
| $M_{neg}$ | 877 | 389 | 1 848 |
| $P_{inc}$ | 13,39% | 19,44% | 17,70% |
| $M_{inc}$ | 1 189 | 714 | 2 769 |

TABLEAU 1 – Statistiques du corpus BioScope

```
<sentence id="S1.7">These results <xcope id="X1.7.2"><cue type="speculation" ref="X1.7.2">indicate that</cue>
in monocytic cell lineage, HIV-1 <xcope id="X1.7.1"><cue type="speculation" ref="X1.7.1">could</cue> mimic
some differentiation/activation stimuli allowing nuclear NF-KB expression</xcope></xcope>.</sentence>
```

FIGURE 1 – Extrait du corpus BioScope

## 3.2 CoNLL-2010

Dans le cadre des tâches communes (shared task) de la *Conference on Computational Natural Language Learning* de 2010 (Farkas *et al.*, 2010), trois jeux de données ont été mis à la disposition des participants. La première tâche consistait à identifier les phrases qui contiennent des informations incertaines ou douteuses à l'aide de textes où seuls les marqueurs sont annotés. Le premier jeu de données est composé de paragraphes issus de Wikipedia où apparaissent des « weasel words », c'est-à-dire, des mots et phrases qui donnent l'impression qu'une déclaration spécifique ou significative a été faite, alors qu'il s'agit en réalité d'une revendication vague et ambiguë. Le deuxième jeu de données est issu de BioScope, mais sans les compte-rendus cliniques. La seconde tâche consistait à identifier la portée de l'incertitude dans les données biomédicales, les marqueurs d'incertitude et leurs portées y sont annotés.

## 3.3 *SEM-2012

Prononcé « *starsem* » (*Conference on Lexical and Computational Semantics : *SEM*) propose chaque année une tâche commune. Celle de 2012 s'est attaquée à deux aspects de la négation : la détection de la portée et de la focalisation qui correspond à la partie de la portée la plus explicitement niée. Dans l'exemple suivant, la portée est entre crochets, l'objet entre * et la focalisation soulignée.

[John had] never [*said* as much before].

L'ensemble des données est composé d'un roman et de trois nouvelles de Sherlock Holmes par Conan Doyle, *The Hound of the Baskervilles* pour l'entraînement, *The Adventures of Wisteria Lodge* pour le développement, et *The Adventure of the Red Circle* et *The Adventure of the Cardboard Box* pour les tests. Toutes les occurrences de la négation sont annotées, ce qui donne 1 056 phrases sur 3 899. Pour chaque occurrence, le marqueur et la portée sont annotés, ainsi que l'objet de la négation, s'il y en a un. Les marqueurs et les portées peuvent être discontinus dans ce corpus. Le guide d'annotation a été publié également (Morante *et al.*, 2011). Dans le tableau 2, nous présentons un exemple d'annotation, avec #C numéro de chapitre, #P numéro de phrase, #U position de l'unité lexicale dans la phrase.

Partant du constat que les descripteurs lexicaux et syntaxiques (*lemme, POS, Chunk*) ont été massivement utilisés dans le cadre de CoNLL-2010, les organisateurs de *SEM-2012 les ont inclus d'office

| #C | #P | #U | Unité | Lemme | POS | Chunk | Marqueur | Portée | Évènement |
|---|---|---|---|---|---|---|---|---|---|
| baskervilles01 | 150 | 0 | “ | “ | “ | (SQ* | _ | _ | _ |
| baskervilles01 | 150 | 1 | Then | Then | RB | (ADVP*) | _ | _ | _ |
| baskervilles01 | 150 | 2 | had | have | VBD | * | _ | had | _ |
| baskervilles01 | 150 | 3 | you | you | PRP | (NP*) | _ | you | _ |
| baskervilles01 | 150 | 4 | not | not | RB | (ADVP* | not | _ | _ |
| baskervilles01 | 150 | 5 | better | well | RB | *) | _ | better | _ |
| baskervilles01 | 150 | 6 | consult | consult | VB | (VP* | _ | consult | _ |
| baskervilles01 | 150 | 7 | him | him | PRP | (NP*)) | _ | him | _ |
| baskervilles01 | 150 | 8 | ? | ? | . | * | _ | _ | _ |
| baskervilles01 | 150 | 9 | ” | ” | ” | *) | _ | _ | _ |

TABLEAU 2 – Extrait – Données d’entrainement *SEM-2012

dans leur corpus. En outre, ces données annotées peuvent aussi aider à la conception de systèmes experts spécifiquement conçus pour l’ensemble de données en question. Dans la section suivante, nous attaquons la description des travaux de recherche sur la détection automatique antérieurs et postérieurs à la mise à disposition de ces données.

# 4 Détection automatique : description des approches

D’une manière générale, il existe deux familles d’approche pour aborder la détection automatique des marqueurs : l’utilisation de systèmes experts, raisonnant à partir de faits et règles connus pour répondre à des questions précises, et la classification par apprentissage automatique, à l’aide de méthodes telles que les champs aléatoires conditionnels (*Conditional Random Fields* ou CRFs), les machines à vecteurs support (SVM) ou les réseaux de neurones.

## 4.1 Système expert

Un système expert est un programme qui raisonne à partir d’informations symboliques et utilise des règles expertes afin d’essayer d’atteindre les performances du niveau d’un expert. Pour des tâches spécifiques, il est possible de construire un système qui raisonne aussi bien que les spécialistes de ces domaines. Partant de cette hypothèse, les premiers systèmes dédiés à la détection de la négation sont proposés sous cette forme. Ainsi, au début des années 2000, Chapman *et al.* (2001) propose l’algorithme *NegEx* qui utilise les expressions régulières pour détecter les phrases négatives et identifier les termes médicaux à l’intérieur de la portée de la négation. La même année, Mutalik *et al.* (2001) propose *Negfinder*. Le système combine un analyseur lexical (lexeur) qui utilise des expressions régulières afin de générer un automate fini et un analyseur syntaxique (parser) qui repose sur un sous-ensemble restreint de la grammaire non contextuelle LALR [3]. Ainsi, *NegFinder* permet d’identifier les concepts impactés par la négation dans les textes médicaux lorsqu’ils sont proches du lemme marquant la négation. Une extension des travaux précédents est développée par (Elkin *et al.*, 2005). Le système utilise le *Mayo Vocabulary Server* afin de récupérer un ensemble de concepts médicaux présents dans les fragments de phrase. Ensuite, des règles expertes assignent à chaque concept un descripteur qui indique si le concept est une déclaration positive, négative ou incertaine.

3. Look-Ahead Left Recursive

Par la suite, les systèmes experts ont majoritairement impliqué les traits syntaxiques. Huang & Lowe (2007) proposent d'automatiser la détection de la négation en combinant les expressions régulières et l'analyse grammaticale. Les négations sont classées sur la base des catégories syntaxiques et sont indiquées dans des arbres syntaxiques. Cette approche hybride permet d'identifier les concepts marqués par la négation dans les rapports de radiologie même s'ils sont situés à distance des termes négatifs. Özgür & Radev (2009) émettent l'hypothèse que l'on peut caractériser la portée d'un marqueur de l'incertitude par sa catégorie grammaticale et la structure syntaxique de la phrase. Ils développent des règles pour détecter la portée d'un marqueur à partir des noeuds dans un arbre syntaxique. Øvrelid *et al.* (2010) construisent un petit ensemble de règles qui définissent la portée de chaque marqueur. Pour développer ces règles, les auteurs exploitent les informations contenues dans le guide d'annotation du corpus BioScope, ainsi qu'une analyse manuelle des données d'entraînement pour trouver les interactions des constructions pour divers types de marqueurs. Kilicoglu & Bergler (2010) proposent une méthodologie pour la détection de l'incertitude essentiellement basée sur des règles, qui repose sur une combinaison d'informations lexicales et syntaxiques. L'information lexicale est représentée par un simple dictionnaire, alors que l'information syntaxique nécessaire est identifiée grâce aux arbres d'analyse de constituants et aux arbres d'analyse des dépendances retournés par le Stanford Parser [4]. Apostolova *et al.* (2011) présentent ScopeFinder, un système expert pour la détection de la portée de la négation et de l'incertitude, dont les règles sont construites à partir des patrons lexico-syntaxiques extraits automatiquement du corpus BioScope. La méthode de construction du système fait qu'il est adaptable aux données de différents domaines.

Les systèmes experts présentés dans cette section arrivent à obtenir de bons résultats sur les données biomédicales pour lesquelles ils sont développés. Cependant, pour être appliqués efficacement, ces systèmes doivent être spécifiquement adaptés aux langues et types de données concernés. Les systèmes par apprentissage artificiel souffrent moins de ce problème.

## 4.2 Apprentissage artificiel

Depuis plusieurs années, les méthodes d'apprentissage artificiel permettant de manipuler de grands volumes de données se sont popularisées en TAL. Pour les tâches de classification que nous étudions, l'apprentissage supervisé est principalement utilisé puisque plusieurs ensembles de données annotées sont disponibles. Les systèmes les plus performants utilisent principalement les SVM, les CRF, ainsi que les réseaux de neurones. D'ailleurs, il nous semble que le premier système pour la détection de l'incertitude exploite l'apprentissage artificiel avec un SVM (Light *et al.*, 2004). Dans ce travail, les auteurs utilisent le système pour sélectionner les phrases spéculatives dans des résumés d'articles de MEDLINE pour démontrer que la détection de l'incertitude peut être efficacement traitée en se basant sur les lemmes qui l'impliquent. Dans la plupart des travaux, les systèmes combinent un module de détection des marqueurs de négation, et un module de détection de la portée de cette négation. Nous les examinons ci-dessous.

### 4.2.1 Détection des marqueurs

Dans cette section, nous présentons les systèmes par apprentissage supervisé les plus performants pour la détection des marqueurs de l'incertitude et de la négation. Dans le cadre des tâches communes de CoNLL-2010, les systèmes utilisent, pour la plupart, le GENIA Tagger afin d'obtenir les lemmes,

4. https://nlp.stanford.edu/software/lex-parser.shtml

l'étiquetage morpho-syntaxique (POS), ainsi que l'analyse syntaxique de surface (*Shallow parsing, chunking*). D'autre part, bien que la tâche commune de *SEM-2012 ne porte pas sur la détection des marqueurs, les systèmes soumis reposent tous sur un sous-système de détection des marqueurs dont la description et les résultats sont donnés. Les résultats sont évalués à l'aide de trois métriques : la précision, qui quantifie la pertinence de l'étiquetage, le rappel, qui quantifie la sensibilité de l'étiquetage, ainsi que la moyenne harmonique de la précision et du rappel noté $F_1$. Nous présentons les résultats des systèmes les plus performants dans le tableau 3 en fin de section.

Le système proposé par Tang *et al.* (2010) est le plus performant sur les deux ensembles de données de CoNLL. Le système classe les unités selon une représentation en BIO (*Begin*, *Inside*, *Outsite*). La détection des marqueurs se fait en trois étapes. La première consiste à pré-traiter les données avec le GENIA Tagger. La seconde opère deux classifications indépendantes et adaptées à l'étiquetage de séquences (CRF++-0.53 [5] et SVM$^{hmm}$ [6]) sur les données pré-traitées. La dernière étape consiste à entraîner un CRF en utilisant les prédictions de l'étape précédente. Les descripteurs utilisés sont : les lemmes et les séquences d'étiquetage morpho-syntaxique de l'incertitude retournés par chaque classifieur, la classification des unités en B, I ou O dont l'étiquetage est identique pour les deux classifieurs, la position de l'unité si elle est la dernière unité du marqueur prédite par chaque classifieur, ainsi que l'ensemble des descripteurs utilisés par l'étape précédente.

Le système de Li *et al.* (2010b), deuxième sur le corpus de Wikipedia et troisième sur les données biomédicales, classe les unités selon une représentation en BIO. Le classifieur est entraîné en utilisant un CRF (CRF++ 0.51) avec les paramètres par défaut. Ils utilisent un algorithme glouton pour sélectionner un meilleur ensemble de descripteurs (unité, lemme, POS, chunk) pour le classifieur selon les résultats obtenus sur les données de développement. En post-traitement, ils utilisent des règles expertes pour classer correctement certains marqueurs oubliés ou incorrectement classés.

Le système de Zhou *et al.* (2010), second sur le corpus de données biomédicales, identifie les marqueurs candidats à partir d'une liste de mots clés puis applique un CRF pour corriger les erreurs d'identification. Les marqueurs récupérés sont utilisés comme descripteurs pour la détection par CRF. Les lemmes sont remplacés par les radicaux.

Extension de leurs travaux précédents, le système proposé par Velldal *et al.* (2012) traite l'ensemble de marqueurs en tant que classe fermée. Ce faisant, ils expliquent que l'on peut réduire considérablement le nombre d'exemples et conséquemment le nombre de descripteurs présentés au classifieur sans perdre en rappel. Le système utilise un SVM appliqué en utilisant de simples descripteurs n-gramme sur les mots et lemmes, à la droite et la gauche des marqueurs candidats.

Read *et al.* (2012) et Lapponi *et al.* (2012) sont deux systèmes proposés par des chercheurs de l'université d'Oslo dans le cadre de *SEM-2012. Leur système de détection des marqueurs est une adaptation de leur précédent système de classification (Velldal *et al.*, 2012). La principale modification est la prise en charge de l'identification des marqueurs de négation affixaux annotés dans les données d'entraînement et de développement. Le SVM utilise des descripteurs qui ciblent spécifiquement les marqueurs affixaux. Le premier descripteur enregistre des n-grammes du début et de la fin de la base à laquelle un affixe s'accroche. Par exemple, dans le cas du marqueur *impossible* les n-grammes {possi, poss, pos, ...} et {sible, ible, ble, ...} sont enregistrés et sont joints avec l'affixe *im* et l'étiquette POS *JJ*. Le second descripteur cherche à imiter l'effet d'une recherche dans un lexique pour les séquences restantes auquel un affixe s'attache aussi, en vérifiant son statut en tant que forme de base

5. https://taku910.github.io/crfpp/
6. https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html

indépendante et son étiquetage POS. La raison derrière ce descripteur est que l'occurrence d'une séquence telle que *un* dans un mot comme *underlying* doit être considérée comme peu probable d'être un marqueur étant donné que la première partie de la séquence restante constitue très rarement le début d'un mot. Les marqueurs multi-mots tels que *neither...nor* ou bien l'expression *on the contrary* sont trop rares pour être détecter par le classifieur. Par conséquent, un petit ensemble de règles expertes est développé afin de les détecter en post-traitement.

Après avoir écarté les affixes marquant la négation, le système de Chowdhury (2012) collecte automatiquement un vocabulaire de toutes les unités lexicales qui ne sont pas des marqueurs de négation et dont la taille dépasse trois caractères dans les données d'entraînement. Ensuite, le système les utilise pour extraire des descripteurs qui pourraient être utiles pour identifier des affixes marqueurs de négation potentiels (**un*able*). Une liste d'expressions exprimant très probablement la négation est créée à partir des données d'entraînement en se basant sur leurs fréquences. La liste contient les termes suivants : *nor, neither, without, nobody, none, nothing, never, not, no, nowhere, non*. Les marqueurs de négation affixaux sont identifiés si l'unité est prédite en tant que marqueur par le classifieur et possède l'un des affixes suivants qui sont collectés à partir des données d'entraînement : *less, un, dis, im, in, non, ir*. Les lemmes sont passés en minuscule dans l'ensemble descripteurs. Un post-traitement supplémentaire est effectué pour annoter certains marqueurs présents dans les données d'entraînement mais oubliés par le classifieur lors de la prédiction sur les données de développement. Ces expressions sont *neither, nobody, save for, save upon, et by no means*. Une vérification spéciale est faite pour l'expression *none the less* qui est marquée comme une non-négation dans les données d'entraînement. Finalement, un modèle de CRF est entraîné sur les descripteurs collectés et utilisé pour prédire les marqueurs de négation sur les données de test.

| Données | Système | Apprentissage | Précision | Rappel | $F_1$ |
|---|---|---|---|---|---|
| Wikipedia | Tang *et al.* (2010) | CRFs+SVM | 63,0 | **25,7** | **36,5** |
| | Li *et al.* (2010b) | CRF | **76,1** | 21,6 | 33,7 |
| Biomédical | Tang *et al.* (2010) | CRFs+SVM | 81,7 | **81,0** | **81,3** |
| | Zhou *et al.* (2010) | CRF | 83,1 | 78,8 | 80,9 |
| | Velldal *et al.* (2012) | SVM | 84,8 | 77,2 | 80,8 |
| | Li *et al.* (2010b) | CRF | **87,4** | 73,4 | 79,8 |
| *SEM-2012 | Read *et al.* (2012) | SVM | 91,42 | 92,80 | 92,10 |
| | Lapponi *et al.* (2012) | SVM | 89,17 | **93,56** | 91,31 |
| | Chowdhury (2012) | CRF | **93,41** | 91,29 | **92,34** |

TABLEAU 3 – Résultats pour la détection des marqueurs pour l'incertitude (CoNLL) et la négation (*SEM-2012) (scores en pourcentage, les meilleurs scores sont en gras)

Pour la détection des marqueurs d'incertitude comme de négation, les différences de performance entre les systèmes les plus efficaces sont ténues. Ils sont même équivalents si l'on prend en considération les légères variations de scores qui peuvent intervenir selon les entraînements. Comme souvent, une haute précision entraîne une chute du rappel et inversement. Les systèmes exploitent de nombreux descripteurs et reposent donc sur les performances des étiqueteurs qui restent imparfaites. En outre, la plupart des auteurs cités notent la sous-représentation de certains marqueurs dans les données d'entraînement rend nécessaire l'application de règles de post-traitement dépendantes de la langue.

### 4.2.2 Détection de la portée

Dans cette section, nous présentons les systèmes par apprentissage supervisé les plus performants pour la détection de la portée de l'incertitude et de la négation. Les résultats obtenus sur le corpus BioScope sont évalués en pourcentage de portées correctement identifiées. Pour *SEM-2012, les résultats sont évalués au niveau des unités lexicales identifiées et des portées correctement identifiées (précision, rappel, $F_1$). Nous présentons les résultats des systèmes les plus performants dans les tableaux 4 et 5 en fin de section.

Li *et al.* (2010a) proposent une approche pour l'apprentissage de la portée de la négation utilisant un système d'analyse sémantique superficielle simplifiée qui combine des règles expertes, l'élagage et la classification binaire par SVM. Étant donné un arbre syntaxique et un prédicat lui appartenant, l'analyse sémantique superficielle reconnaît et assigne à tous les constituants dans la phrase le rôle sémantique leur correspondant. Le marqueur de négation peut être considéré comme le prédicat, alors que les constituants de la portée peuvent être considérés comme les arguments du marqueur de la négation. Deux règles expertes sont adoptées pour assigner le rôle d'argument de la portée sur plusieurs constituants dans un arbre syntaxique donné. Excepté le marqueur de négation et ses ancêtres, tous les constituants dans l'arbre syntaxique dont le parent couvre le marqueur sont collectés comme arguments candidats. Finalement, un classifieur binaire [7] est appliqué pour déterminer si les arguments candidats sont valides.

Le système proposé par Velldal *et al.* (2012) combine deux approches. L'une utilise des règles expertes opérant sur des arbres syntaxiques de dépendance. La seconde utilise un SVM d'ordonnancement (*ranking*) qui opère sur des noeuds d'arbres syntaxiques des constituants syntagmatiques.

L'année dernière, un système basé sur un réseau de neurones convolutionnel (CNN) a été proposé par Qian *et al.* (2016). Les descripteurs pertinents sont récupérés à partir des chemins syntaxiques entre les marqueurs et les unités candidates dans les arbres syntaxiques de dépendances et les arbres syntaxiques de constituants syntagmatiques. La couche convolutionnelle concatène ces descripteurs avec leurs positions relatives en un seul vecteur descripteur qui alimente une couche *softmax*. Cette couche cachée calcule les scores de confiance de l'étiquetage pour une classification en OBA où sont classés : O les unités lexicales ne faisant pas parties de la portée, B les unités lexicales de la portée qui apparaissent avant le marqueur, et A les unités lexicales de la portée qui apparaissent après le marqueur.

Dans la continuité de Velldal *et al.* (2012), Read *et al.* (2012) combine un système expert et l'ordonnancement par SVM. Les règles expertes y sont formulées en fonction de patrons cohérents relevés en alignant constituants syntaxiques et les portées. Les règles s'appuient sur la fréquence des chemins allant du marqueur jusqu'au constituant aligné avec la portée sur la base des annotations dans les données d'entraînement (SEM-2012). L'approche motivée par les données pour la résolution de la portée implique l'apprentissage d'une fonction de rang sur les constituants syntaxiques candidats. Cela commence par la sélection de négations pour lesquelles l'arbre syntaxique contient un constituant qui s'aligne avec la portée. Ensuite, la sélection d'un candidat initial est effectuée en prenant le constituant le plus petit qui s'étend sur tous les mots dans la portée. Finalement, la génération d'un candidat subséquent est effectuée en traversant le chemin vers la racine de l'arbre. Les candidats dont la projection correspond à la portée sont étiquetés comme corrects ; tous les autres sont étiquetés comme incorrects. Expérimentant avec divers types de descripteurs, ils utilisent la fonction d'ordonnancement ordinal présente dans le SVM-light toolkit pour apprendre une fonction

7. http://svmlight.joachims.org/

score linéaire pour choisir les portées candidates correctes.

Le système de détection de la portée de (Lapponi *et al.*, 2012) utilise l'implémentation CRF Wapiti [8] avec une version spéciale de la séquence d'étiquetage IOB, où les marqueurs de négation sont étiquetés comme démarreurs de séquence et les unités de la portée comme constituants phrastiques (*chunk*). En plus des descripteurs fournis pour chaque unité dans le corpus (lemme, POS, constituant phrastique), étiquettes, la distance entre chaque unité et le marqueur le plus proche à droite et à gauche est extraite. Les descripteurs ont été étendus pour inclure des bigrammes et trigrammes vers l'avant et l'arrière des unités et au niveau des étiquettes POS, ainsi que les unigrammes et bigrammes de POS lexicalisés (couple forme/POS).Les représentations syntaxiques sont converties en représentations des dépendances en utilisant le Stanford Parser. Les descripteurs récupérés à partir des graphes de dépendances ont pour but de modéliser les relations syntaxiques entre chaque unité et le marqueur le plus proche. Finalement, un algorithme basé sur des règles expertes est appliqué en sortie du CRF dans le but d'assigner chaque unité dans la portée à son marqueur et déterminer les chevauchements de portées.

La méthode proposée par (Packard *et al.*, 2014) repose sur deux approches. La première est basée sur un indexeur par *minimal Recursion semantics* (MRS). La MRS est un cadre applicatif pour la sémantique computationnelle. Il peut être implémenté selon des formalismes tels que la grammaire syntagmatique guidée par les têtes et la grammaire lexicale-fonctionnelle. La seconde approche reprend le système de (Read *et al.*, 2012) et intervient pour la détection des marqueurs et pour la détection de la portée dans les cas où le *LinGo English Ressource Grammar* (ERG) ne peut pas créer de représentation sémantique fiable pour une phrase. La MRS rend explicite les relations prédicat-argument, ainsi que des informations partielles concernant la portée. Un système expert traverse les structures MRS des phrases négatives.

Fancellu *et al.* (2016) proposent deux approches par réseau de neurones pour résoudre le problème de détection de la portée de la négation. La première utilise un réseau de neurones *Feed-forward*, c'est-à-dire, un réseau de neurones artificiel où les connexions entre les unités ne forment pas de boucle. Dans ce type de réseau, l'information se déplace uniquement dans un sens, à partir des noeuds de la couche entrée, à travers les noeuds de la couche cachée et finalement des noeuds de sortie. La seconde approche, qui est la plus efficace, utilise un réseau de neurones Long Short-Term Memory (LSTM) bidirectionnel. Un LSTM bidirectionnel est la combinaison de deux structures de réseau de neurones récurrents (RNN). Le RNN bidirectionnel proposé par (Schuster & Paliwal, 1997), et le LSTM proposé par (Hochreiter & Schmidhuber, 1997). Ce réseau est motivé par le fait qu'un être humain raisonne en s'appuyant sur les connaissances qu'il a acquises et qui restent dans la mémoire. Si les réseaux de neurones « classiques » ne sont pas capables de copier ce processus, les RNN répondent à ce problème. En effet, ce sont des réseaux qui bouclent, ce qui permet aux informations de persister. Parmi les RNN, les réseaux LSTM sont les plus efficaces pour apprendre des dépendances de long-terme et sont donc plus à même de résoudre, par exemple, le problème de la portée discontinue. Le LSTM de (Fancellu *et al.*, 2016) bidirectionnel est un modèle séquentiel qui opère dans le sens de lecture et dans le sens contraire. La passe arrière est particulièrement importante dans le cas de la détection de la portée puisque les unités affectées peuvent se trouver avant le marqueur. Par ailleurs, les cellules LSTM sont plus efficaces pour retenir les informations utiles lors de la rétropropagation du gradient, qui permet de corriger les différences entre les prédictions sortantes et celles désirées en calculant le gradient de l'erreur pour chaque neurone, de la dernière couche vers la première. Le système ne s'entraîne que sur les phrases porteuses d'une négation. Le système de base prend

8. https://wapiti.limsi.fr/

| | Système | Apprentissage | Résumé | Clinique | Article |
|---|---|---|---|---|---|
| Incertitude | Velldal *et al.* (2012) | SVM | 79,56 | **78,69** | **75,15** |
| | Qian *et al.* (2016) | CNN | **85,75** | 73,92 | 59,82 |
| Négation | Li *et al.* (2010a) | SVM | **81,84** | 89,79 | 64,02 |
| | Velldal *et al.* (2012) | SVM | 74,35 | **90,74** | **70,21** |
| | Qian *et al.* (2016) | CNN | 77,14 | 89,66 | 55,32 |

TABLEAU 4 – Comparaison des systèmes les plus performants sur le corpus BioScope ; les résultats sont donnés en pourcentage de portées correctement identifiées (les meilleurs scores sont en gras)

| | Système | Apprentissage | Précision | Rappel | $F_1$ |
|---|---|---|---|---|---|
| ULI | Read *et al.* (2012) | SVM | 81,99 | 88,81 | 85,26 |
| | Lapponi *et al.* (2012) | CRF | 86,03 | 81,55 | 83,73 |
| | Packard *et al.* (2014) | SVM | 86,1 | **90,4** | 88,2 |
| | (Fancellu *et al.*, 2016) | BiLSTM | **92,62** | 85,13 | **88,72** |
| PCI | Read *et al.* (2012) | SVM | 87,43 | 61,45 | 72,17 |
| | Lapponi *et al.* (2012) | CRF | 85,71 | 62,65 | 72,39 |
| | Packard *et al.* (2014) | SVM | 98,8 | **65,5** | **78,7** |
| | Fancellu *et al.* (2016) | BiLSTM | **99,40** | 63,87 | 77,77 |

TABLEAU 5 – Comparaison des systèmes les plus performants sur les données de test de *SEM-2012, évalués au niveau des unités lexicales identifiées (ULI) et des portées correctement identifiées (PCI) ; les résultats sont donnés en pourcentage (les meilleurs scores sont en gras)

en entrée une instance $I(n, c)$ ou chaque mot est représenté par un vecteur $n$ (*word-embedding*) et par un vecteur $c$ qui détermine si le mot fait parti d'un marqueur (*cue-embedding*). Dans le but d'affiner la prédiction, le système peut aussi utiliser une représentation vectorielle de l'étiquetage morpho-syntaxique pour chaque mot.

Tous les systèmes de détection de la portée par CRF et SVM que nous avons présentés sont dépendants des performances de leurs systèmes de détection des marqueurs. La multiplication des descripteurs de plus en plus complexes rendent les systèmes difficiles à adapter pour une autre langue. Les descripteurs complexes et dépendants de la langues tels que la MRS (Packard *et al.*, 2014) ou encore les chemins syntaxiques (Qian *et al.*, 2016) dépendent d'outils externes dont les performances sont imparfaites. Le système de Fancellu *et al.* (2016) nous semble être le plus prometteur car il est facilement portable d'un langage à l'autre du moment que des données annotées sont disponibles.

### 4.2.3 Expériences & résultats

Le code de Fancellu *et al.* (2016) étant disponible en ligne [9], nous avons pu mettre à l'épreuve son système et procéder à quelques changements. Nous avons d'abord mis à jour le code vers la version 1.0 de Tensorflow [10], puis nous l'avons exécuté sur notre système. Bien que nous nous attendions à de légères différences entre les entraînements, nous constatons une instabilité dans les résultats obtenus. Les scores $F_1$ présentés dans le tableau 6 illustrent ces variations, qui atteignent jusqu'à 6 points d'écart. Il est fort probable que le faible volume de données d'entraînement (848 phrases

9. https://github.com/ffancellu/NegNN
10. https://www.tensorflow.org/

| Apprentissage | Unités de la portée | | | Portées exactes | | |
|---|---|---|---|---|---|---|
| | $F_1min$ | $F_1max$ | $F_1moy$ | $F_1min$ | $F_1max$ | $F_1moy$ |
| BiLSTM | 88,20 | **89,68** | **88,83** | 72,46 | **78,34** | **76,00** |
| BiGRU | 87,28 | 89,58 | 88,39 | 73,08 | 77,49 | 74,87 |

TABLEAU 6 – Variation des scores $F_1$ du BiLSTM de (Fancellu *et al.*, 2016) et de notre BiGRU (les scores sont en pourcentage, les meilleurs scores sont en gras)

négatives) cause cette instabilité. D'autre part, nous avons remplacé les cellules LSTM du réseau par des cellules GRU (Gated Reccurent Unit – Cho *et al.* (2014)) qui, en fusionnant plusieurs étapes du LSTM, permettent un calcul simplifié et plus rapide (40% de gains sur notre système). L'ensemble de notre code est accessible sur `https://github.com/Kureman/NegNN`.

# 5 Conclusions

Entre 2008 et 2012, les travaux sur la détection automatique de la négation et de l'incertitude en langue anglaise se sont multipliés grâce à la sortie de BioScope ainsi que les tâches communes de CoNLL-2010 et *SEM-2012. Depuis, plusieurs travaux innovants, principalement axés sur la détection de la portée, ont vu le jour.

Dans cet article, nous avons d'abord présenté les marqueurs ainsi que les différents types de portées posant problème et que l'on peut rencontrer dans les données dont nous disposons. Nous avons ensuite présenté les corpus annotés dont les chercheurs ont disposé pour entraîner des modèles par apprentissage automatique, pour extraire des descripteurs supplémentaires, ou bien pour la conception de systèmes experts. La revue de ces différents travaux nous a montré que ces systèmes experts sont désormais délaissés au profit des approches par apprentissage artificiel, les plus performants dans les évaluations dont nous avons rendu compte. C'est donc sur ces derniers systèmes que nous nous sommes penchés, en tentant notamment de reproduire les résultats de Fancellu *et al.* (2016) ; nous avons remarqué une instabilité du système, due au manque de données, qui souligne une limite de ce type d'approche. Nous avons aussi proposé une modification de son approche pour obtenir des résultats équivalents mais avec un temps de calcul bien plus court.

Les travaux présentés dans cet article constituent un point de départ pour nos prochains travaux qui nécessiteront l'annotation de données en français. Nous pourrons notamment nous inspirer des consignes d'annotation disponibles pour l'anglais et du volume de données disponible dans le but d'obtenir un ensemble de données suffisamment large et varié pour stabiliser un modèle.

# Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

# Références


APOSTOLOVA E., TOMURO N. & DEMNER-FUSHMAN D. (2011). *Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes*, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*. Association for Computational Linguistics.

CHAPMAN W. W., BRIDEWELL W., HANBURY P., COOPER G. F. & BUCHANAN B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, **34**(5).

CHO K., VAN MERRIËNBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*.

CHOWDHURY M. F. M. (2012). *FBK : Exploiting phrasal and contextual clues for negation scope detection*, In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

ELKIN P. L., BROWN S. H., BAUER B. A., HUSSER C. S., CARRUTH W., BERGSTROM L. R. & WAHNER-ROEDLER D. L. (2005). A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, **5**.

FANCELLU F., LOPEZ A. & WEBBER B. (2016). *Neural networks for negation scope detection*, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1.

FARKAS R., VINCZE V., MÓRA G., CSIRIK J. & SZARVAS G. (2010). *The CoNLL-2010 shared task : learning to detect hedges and their scope in natural language text*, In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural Comput.*, **9**(8).

HUANG Y. & LOWE H. J. (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association : JAMIA*, **14**(3).

KILICOGLU H. & BERGLER S. (2010). *A High-Precision Approach to Detecting Hedges and their Scopes.*, In *CoNLL Shared Task*.

LAPPONI E., VELLDAL E., ØVRELID L. & READ J. (2012). *Uio 2 : sequence-labeling negation using dependency features*, In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

LI J., ZHOU G., WANG H. & ZHU Q. (2010a). *Learning the scope of negation via shallow semantic parsing*, In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 671–679. Association for Computational Linguistics.

LI X., SHEN J., GAO X. & WANG X. (2010b). *Exploiting rich features for detecting hedges and their scope*, In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics.

LIGHT M., QIU X. Y. & SRINIVASAN P. (2004). *The language of bioscience : Facts, speculations, and statements in between*, In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases : tools for users*. Association for Computational Linguistics.

MORANTE R. & BLANCO E. (2012). ** SEM 2012 shared task : Resolving the scope and focus of negation*, In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

MORANTE R., SCHRAUWEN S. & DAELEMANS W. (2011). Annotation of negation cues and their scope. guidelines v1.0.

MUTALIK P. G., DESHPANDE A. & NADKARNI P. M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents : a quantitative study using the umls. *Journal of the American Medical Informatics Association : JAMIA*, **8**(6).

PACKARD W., BENDER E. M., READ J., OEPEN S. & DRIDAN R. (2014). *Simple Negation Scope Resolution through Deep Parsing : A Semantic Solution to a Semantic Problem.*, In *ACL (1)*.

QIAN Z., LI P., ZHU Q., ZHOU G., LUO Z. & LUO W. (2016). : Association for Computational Linguistics.

READ J., VELLDAL E., ØVRELID L. & OEPEN S. (2012). *Uio 1 : Constituent-based discriminative ranking for negation resolution*, In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

SCHUSTER M. & PALIWAL K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, **45**(11).

TANG B., WANG X., WANG X., YUAN B. & FAN S. (2010). *A cascade method for detecting hedges and their scope in natural language text*, In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics.

VELLDAL E., ØVRELID L., READ J. & OEPEN S. (2012). Speculation and negation : Rules, rankers, and the role of syntax. *Computational Linguistics*, **38**(2).

VINCZE V., SZARVAS G., FARKAS R., MÓRA G. & CSIRIK J. (2008). The bioscope corpus : biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, **9**(Suppl 11).

ZHOU H., LI X., HUANG D., LI Z. & YANG Y. (2010). *Exploiting multi-features to detect hedges and their scope in biomedical texts*, In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics.

ÖZGÜR A. & RADEV D. R. (2009). *Detecting speculations and their scopes in scientific text*, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3-Volume 3*. Association for Computational Linguistics.

ØVRELID L., VELLDAL E. & OEPEN S. (2010). *Syntactic scope resolution in uncertainty analysis*, In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics.

# Normalisation de termes complexes par sémantique distributionnelle guidée par une ontologie

Arnaud Ferré[1, 2]
(1) MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
(2) LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
arnaud.ferre@universite-paris-saclay.fr

## RESUME

Nous proposons dans cet article une méthode semi-supervisée originale pour la création de représentations vectorielles pour des termes (complexes ou non) dans un espace sémantique pertinent pour une tâche de normalisation de termes désignant des entités dans un corpus. Notre méthode s'appuie en partie sur une approche de sémantique distributionnelle, celle-ci générant des vecteurs initiaux pour chacun des termes extraits. Ces vecteurs sont alors plongés dans un autre espace vectoriel construit à partir de la structure d'une ontologie. Pour la construction de ce second espace vectoriel ontologique, plusieurs méthodes sont testées et comparées. Le plongement s'effectue par entraînement d'un modèle linéaire. Un calcul de distance (en utilisant la similarité cosinus) est enfin effectué pour déterminer la proximité entre vecteurs de termes et vecteurs de concepts de l'ontologie servant à la normalisation. La performance de cette méthode a atteint un rang honorable, ouvrant d'encourageantes perspectives.

## ABSTRACT

**Normalization of complex terms with distributional semantics guided by an ontology**

We propose in this paper an original and semi-supervised method for computing continuous vector representations for terms (complex or non-complex) in a semantically structured space relevant to a task of normalization of terms of designating entities in a corpus. Our method partially based on a distributional semantics approach, which generates initial vectors for each of the extracted terms. Then, these vectors are embedded into another vector space constructed from the structure of the ontology. For the construction of this second ontological vector space, several methods are tested and compared. This embedding is carried out by training a multivariate linear regression. Finally, we apply a distance calculation (using cosine similarity) to determine the proximity between vector of terms and vector of concepts used for the normalization. The performance of this method reaches an honorable rank, opening up encouraging perspectives.

MOTS-CLES : TAL, extraction d'information, étiquetage par une ontologie, espace vectoriel, sémantique distributionnelle, modèle linéaire

KEYWORDS: NLP, information extraction, ontology-based tagging, vector space, distributional semantics, multivariate linear regression

# 1 Introduction

La normalisation consiste à annoter des mots ou des combinaisons de mots avec une ou plusieurs catégories sémantiques. La normalisation rencontre plusieurs difficultés, comme la variabilité importante de la morphologie des termes, qu'ils soient représentés par un mot ou par plusieurs (Nazarenko et al., 2006). Les termes multi-mots présentant des structures morphosyntaxiques variées et des imbrications complexes, tels que les groupes nominaux complexes (*complex noun phrase*) sont particulièrement difficiles à étiqueter par des catégories. En effet, la morphologie des labels des catégories sémantiques n'est pas nécessairement proche de la morphologie des termes à annoter. Or, dans les textes de la littérature spécialisée, les groupes nominaux complexes se retrouvent en abondance (Maniez, 2007). Une approche par similarité morphologique entre terme et étiquette sémantique apparaît donc limitée pour effectuer cette tâche (Golik et al., 2011). Une autre difficulté vient du nombre important de catégories sémantiques souvent à utiliser, rendant une approche par classification supervisée coûteuse en annotation manuelle (e.g. plus de trois millions de concepts dans le metathesaurus biomédical de l'UMLS de catégories pour le thésaurus biomédical UMLS (McCray, 1989)).

Une ontologie informatique est une représentation formelle et partielle de propriétés générales d'un ensemble de connaissances, et représente donc un objet manipulable par un programme informatique. Dans le cas de cette article, nous n'emprunterons aux ontologies que la notion de concept (e.g. un concept labélisé 'Chien' qui représente l'abstraction de tous les animaux domestiques de l'espèce *Canis lupus*) et celle de relation hiérarchique (e.g. le concept 'Labrador' hérite du concept 'Chien' qui hérite lui-même du concept 'Animal', etc.). Les concepts d'une ontologie peuvent alors être utilisés pour représenter des catégories sémantiques de façon formelle et structurée.

Une alternative aux approches par classification supervisée ou aux approches morphologiques consiste à calculer la proximité sémantique entre des termes par sémantique distributionnelle. C'est une approche fondée sur la corrélation entre la similarité de sens et la similarité de distribution des unités sémantiques (mot, combinaison de mots, phrase, documents, …) (Firth, 1957; Harris, 1954). Une unité sémantique peut être représentée par un vecteur construit à partir de la distribution des informations de contexte dans lesquels elle est trouvée. La proximité des vecteurs dans cet espace est alors transposable à une proximité sémantique (Fabre et al., 2015). Il existe aujourd'hui de nombreuses méthodes de génération de tels espaces vectoriels, tel que Word2Vec (Mikolov et al., 2013), mais celles-ci se concentrent habituellement sur les jeux de données massifs (Fabre et al., 2014) dans lesquels l'information est relativement répétée.

La question qui nous intéresse ici est : comment utiliser la sémantique distributionnelle pour normaliser les termes par une ontologie ? C'est-à-dire comment relier l'information distributionnelle aux concepts d'une ontologie ? Et comment répondre à cette question pour des relativement petits corpus et un grand nombre de concepts ? Cet article propose une méthode s'appuyant sur la génération de vecteurs dans un espace vectoriel des termes (EVT) ainsi que sur l'étude et la comparaison de vecteurs générés pour former un espace vectoriel des concepts de l'ontologie (EVO), puis sur le plongement de l'EVT dans l'EVO.

# 2 Matériel

Les données utilisées sont celles de la tâche de catégorisation Bacteria Biotope (tâche 3) de BioNLP Shared Task en 2016 (Deléger et al., 2016). Les documents sont des références de PubMed, composées de titres et de résumés d'articles scientifiques dans le domaine de la biologie. La tâche consiste, étant donné les entités du corpus dénotant les habitats bactériens (i.e. des termes extraits), à leur assigner un concept de l'ontologie. Le corpus Bacteria Biotope est divisé en trois : le corpus d'entraînement, le corpus de développement et le corpus de test. Dans le corpus d'entraînement et le corpus de développement les concepts des entités sont données : elles nous ont servi à entraîner notre méthode. Le corpus de test est celui pour lequel les concepts sont à prédire : il nous servira à évaluer notre méthode pour la tâche de normalisation. Chacun de ces corpus a été annoté manuellement. Voici un résumé de leurs caractéristiques :

| | **BB** | | | |
|---|---|---|---|---|
| | Entr. | Dév. | Test | **Total** |
| Documents | 71 | 36 | 54 | **161** |
| Mots | 16 295 | 8 890 | 13 797 | **38 982** |
| Entités | 747 | 454 | 720 | **1 921** |
| Entités distinctes | 476 | 267 | 478 | **1 125** |
| Concepts | 825 | 535 | 861 | **2 221** |
| Concepts distincts | 210 | 122 | 177 | **329** |

TABLE 1 : Statistiques descriptives du corpus BB

En plus de ce corpus, un corpus élargi du même domaine a été utilisé pour générer des représentations vectorielles de chaque mot. Il est composé de 100 000 phrases venant de titres et de résumés d'articles scientifiques dans le domaine de la biologie disponibles sur PubMed. Cela représente un corpus de taille relativement moyenne, qui contient également une majorité de mots non-outils avec une faible fréquence d'apparition (cf. TABLE 2).

| | |
|---|---|
| Nombre mots uniques non-outils | 209 345 |
| Nombre de mots | 4 826 058 |
| Nombre de phrases | 100 250 |

TABLE 2 : Statistiques descriptives du corpus élargi

# 3 Méthode

La méthode présentée dans cet article pour normaliser des termes d'un corpus peut se décomposer en 3 étapes principales : la première étape consiste à générer un espace vectoriel dans lequel les termes du corpus sont représentés (i.e. par des vecteurs) à partir d'une méthode de sémantique distributionnelle ; la seconde étape consiste à générer un autre espace vectoriel dans lequel les concepts de l'ontologie sont représentés ; enfin la troisième étape consiste à déterminer une transformation linéaire permettant de projeter les vecteurs de termes du corpus d'entraînement issu

du premier espace vectoriel dans le second en cherchant à minimiser la distance entre les projections et les vecteurs de concepts associés aux termes, permettant ainsi de prédire des projections du corpus de test et de proposer pour chaque terme un concept en prenant le concept le plus proche dans l'espace (cf. FIGURE 1 et FIGURE 2).

### 3.1 Génération de l'espace vectoriel des termes (EVT)

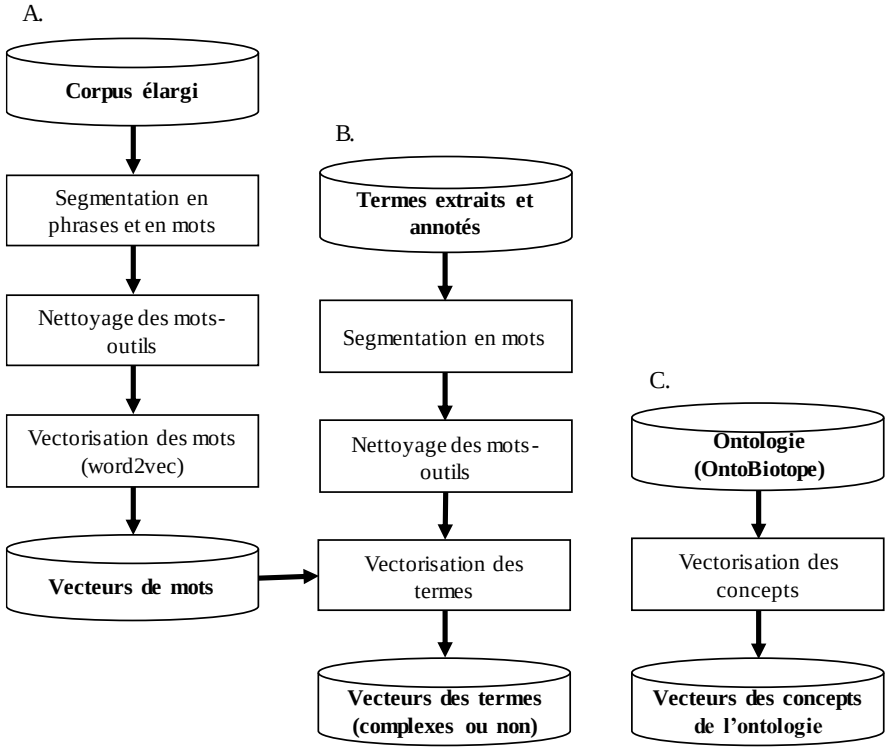



FIGURE 1 : A. Processus de création de vecteurs de mots. B. Processus de création des vecteurs de termes C. Processus de création de vecteurs de concepts

L'espace vectoriel des termes (EVT) est obtenu en générant un vecteur pour chacun des mots du corpus élargi qui comprend également les corpus d'entraînement et de développement. Pour cela, nous avons utilisé l'outil Word2Vec (Mikolov et al., 2013) en prenant pour contexte d'un mot, tous les mots contenus dans la phrase. Pour avoir suffisamment de données à entraîner pour la génération de vecteurs de mots, et aussi pour éviter de prendre en compte des fautes de frappes ou des erreurs, il est habituellement conseillé d'utiliser Word2Vec en ne prenant pas en compte les mots n'apparaissant qu'une ou deux fois dans tout le corpus. Notre corpus contenant beaucoup de mots d'intérêt à faible fréquence, nous avons fait le choix de ne pas appliquer ce seuil de fréquence. Après quelques tests de performance, une dimension de 200 a été choisie pour les vecteurs de sorties (cf. FIGURE 1A), ce qui est du même ordre que ce qui est conseillé habituellement (Mikolov et al., 2013). Pour calculer des représentations vectorielles pour des termes extraits (cf. FIGURE 1B), on peut alors commencer par les segmenter en mots. Pour chaque mot non-outil, on récupère le vecteur issu du calcul précédent. Puis on calcule le vecteur moyen :

$$v_{t_k} = \sum_{i=1}^{n_k} v_{m_i^k}/n_k \qquad (1)$$

Où $v_{t_k}$ est le vecteur associé au terme $t_k$, $n_k$ est le nombre de mots non-outils du terme $t_k$, $v_{m_i^k}$ est le vecteur du mot $m_i^k$ issu de Word2Vec, et le terme $t_k$ est tel que : $\forall i \in [1, n_k], m_i^k \in t_k$

### 3.2 Génération de l'espace vectoriel ontologique (EVO)

La caractéristique commune à tous les différents EVO que nous présentons ici est qu'ils ont une dimension égale au nombre de concepts différents dans l'ontologie. Pour construire les vecteurs de concepts, on initialise des vecteurs nuls possédant autant de dimension que de concepts dans l'ontologie. Chaque valeur du vecteur correspond donc à un des concepts de l'ontologie. On a donc :

$$\forall k \in [\![1, n]\!], v_{c_k} = \left( w_{c_k}^0, \dots, w_{c_k}^i, \dots, w_{c_k}^n \right) \qquad (2)$$

où $v_{c_k}$ est le vecteur associé au concept $c_k$, $c_k$ est le concept associé à la dimension k, n est le nombre de concepts dans l'ontologie et $w_{c_k}^i$ est la valeur du vecteur $v_{c_k}$ pour la dimension i.

Nous présentons dans cet article quatre méthodes différentes pour construire les vecteurs de concepts :

1. La méthode « One-Hot » : Chaque concept est représenté par un vecteur différent, mais aucune information ne permet de représenter la structure de l'ontologie. Ici, seule la valeur associée au concept courant est non-nulle. En utilisant l'équation (2), on a donc :
$$\forall i, k \in [\![1, n]\!], w_{c_k}^i = \begin{cases} 1 \; si \; i = k \\ 0 \; sinon \end{cases} \qquad (3)$$

2. La méthode « Ancestry » : Chaque vecteur de concept possède l'information des vecteurs qui sont ses ancêtres, mais pas l'information de l'organisation de sa lignée. Pour cela, en plus de la valeur associée au concept qui est mise à 1 comme dans la méthode précédente, on va également mettre à 1 tous les parents (directs ou non) du concept courant jusqu'à la racine. En utilisant l'équation (2), on a donc :
$$\forall i, k \in [\![1, n]\!], w_{c_k}^i = \begin{cases} 1 & si \; i = k \\ 1 & si \; c_i \; ancêtre \; de \; c_k \\ 0 & sinon \end{cases} \qquad (4)$$

3. La méthode « Ancestry+ » : S'appuyant sur la méthode précédente, elle vise à éloigner chaque vecteur de concept des vecteurs de ses concepts fils et à l'inverse de le rapprocher des vecteurs de ses concepts parents. L'objectif de cette modification est que, pour un concept devant annoter correctement un terme, il est préférable d'annoter par un concept parent plutôt que par un concept fils (e.g. si le terme 'le chien du voisin' échoue à être annoté par un concept 'Chien', il vaut mieux qu'il soit annoté par un concept 'Animal' que 'Labrador'). D'après les critères d'évaluation de la tâche de normalisation, cette caractéristique pourrait donc avoir une influence positive. Pour cela, au lieu de mettre la valeur associée au concept à 1, celle-ci va prendre une valeur égale à la distance du concept de la racine. Si on définit la fonction $dist$ suivante :
$$\forall i, j \in [\![1, n]\!], dist(c_i, c_j) = distance \; de \; c_i \; à \; c_j \qquad (5)$$
$$Telle \; que : \forall k \in [\![1, n]\!], dist(c_k, c_k) = 0 \; et \; dist(c_k, parent \; de \; c_k) = 1 \qquad (6)$$

   On a alors :

$$w_{c_k}^i = \begin{cases} dist(c_k, racine) + 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases} \quad (7)$$

4. La méthode « Centered Node » : Cette méthode, proposée dans (Eidoon et al., 2007) permet d'augmenter l'information portée par chaque vecteur sur la structure de l'ontologie. Plus précisément, chaque vecteur de concept porte l'information de sa distance à chaque concept de sa lignée (ses ancêtres comme ses descendants). Pour cela, la valeur du concept courant est là aussi mise à 1. De façon symétrique (i.e. vers la racine ou vers les feuilles), plus l'on va s'éloigner de la valeur du concept courant en restant sur la lignée du concept, plus ses valeurs associées aux fils et aux parents vont diminuer. En utilisant la fonction définie en (5) :

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1/(dist(c_i, c_k) + 1) & si\ c_i\ ancêtre\ ou\ descendant\ de\ c_k \\ 0 & sinon \end{cases} \quad (8)$$

## 3.3 Entraînement et prédiction

L'objectif de la phase d'entraînement est de déterminer une transformation de l'EVT vers l'EVO qui minimise toutes les distances entre les vecteurs des termes résultants dans l'EVO et les vecteurs des concepts associés. Dans cet article, une transformation linéaire est étudiée car nous faisons l'hypothèse qu'il y a une certaine similitude de répartition entre les vecteurs de termes dans l'EVT et les vecteurs de concepts associés dans l'EVO. Autrement dit, une transformation non-linéaire pourrait fortement déformer la répartition des vecteurs de termes dans l'EVO pour s'adapter aux données d'entraînement peu nombreuses et ne recouvrant qu'une faible partie des annotations pouvant être détectées dans les textes ciblés. Cet entraînement vise à obtenir les meilleurs paramètres pour approximer l'équation matricielle suivante :

$$Y = X.B + U \quad (9)$$

où Y est une matrice formée d'une série de vecteurs de concepts, X est une matrice formée d'une série de vecteurs de termes (où la ième ligne de X représente le vecteur d'un terme qui est annoté par un concept qui a pour vecteur la ième ligne de Y), B est la matrice contenant les paramètres qui sont à estimer, et U est une matrice contenant une distribution gaussienne multivariée. Cet entraînement est réalisé sur les corpus d'entraînement et de développement (cf. FIGURE 2A).

La matrice obtenue nous permet de concevoir une fonction de transformation linéaire, afin de permettre de prédire de nouveaux vecteurs associés aux termes du corpus d'essai exprimé dans l'EVO :

$$f: \begin{pmatrix} EVT \rightarrow EVO \\ v_{\text{term}} \ \rightarrow \ v'_{\text{term}} = f(v_{\text{term}}) \end{pmatrix} \quad (10)$$

Où $v_{\text{term}}$ est un vecteur de terme dans l'EVT et $v'_{\text{term}}$ est le vecteur résultant du même terme projeté dans l'EVO. Pour satisfaire aux exigences de la tâche d'évaluation, le vecteur de concept le plus proche (en terme de similarité cosinus) de $v'_{\text{term}}$ est choisi pour le terme annoté (cf. FIGURE 2B).

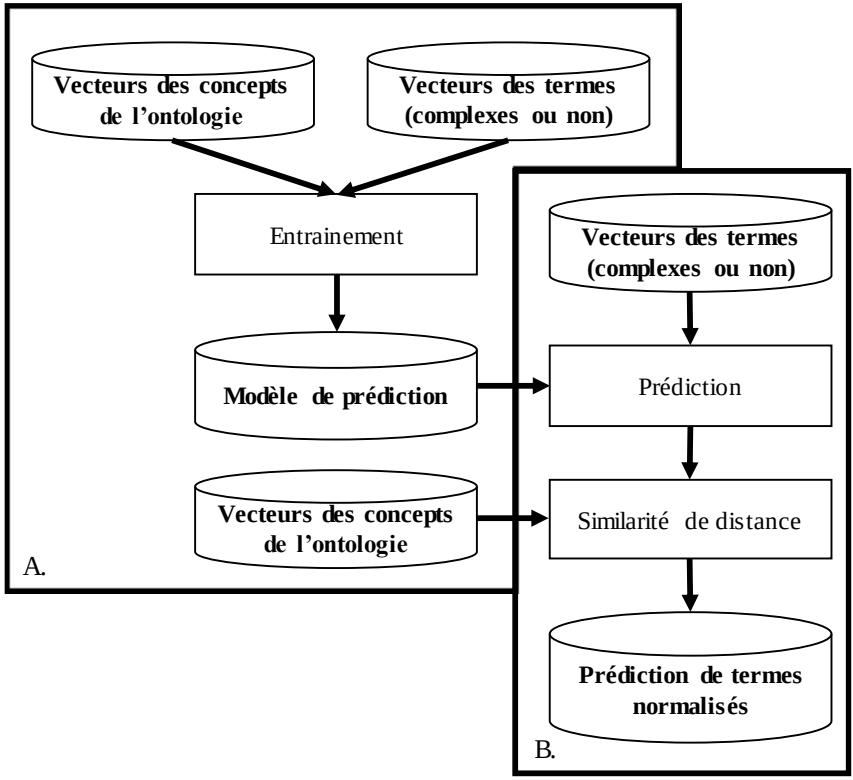



FIGURE 2 : A. Processus d'entraînement pour déterminer une transformation de l'EVT vers l'EVO ; B. Processus de prédiction des concepts associées aux termes extraits

# 4 Résultats

Dans cette section, nous présentons les résultats intermédiaires obtenus pour les deux espaces vectoriels générés, ainsi que les résultats finaux pour la tâche de normalisation en fonction de la méthode de génération de l'EVO utilisée.

## 4.1 Espace vectoriel des termes (EVT) généré

En dépit de la faible fréquence d'apparition des mots du corpus élargi (cf. TABLE 2), les vecteurs de mots obtenus semblent présenter des proximités relativement cohérente du point de vue de la similarité sémantique des termes associés. De plus, la méthode utilisée pour former des vecteurs pour les termes complexes semblent elle aussi pouvoir conserver cette cohérence. Les cas où la tête sémantique est commune dans ces termes semble permettre de retrouver une similarité cohérente (cf. TABLE 3).

| cell | similarité |
|---|---|
| HCE cell | 0,9999 |
| 13C-labeled cell | 0,9998 |
| parietal cell | 0,9989 |
| Schwann cell | 0,9965 |
| CD8+ T cell | 0,9770 |
| PMN cell | 0,9669 |
| macrophage cell | 0,9473 |
| J774 cell line | 0,9230 |
| eukaryotic cell | 0,9114 |
| macrophage-like cell line J774 | 0,9086 |

TABLE 3 : Termes les plus proches du terme 'cell' dans l'EVT

Il semble également que des différences morphologiques n'empêchent pas l'agglomération de vecteurs de termes proches (cf. TABLE 4), ce qui était une des propriétés désirées.

| younger ones | similarité |
|---|---|
| children less than five years of age | 0,8087 |
| children less than 2 years of age | 0,8060 |
| children less than two years of age | 0,7995 |

TABLE 4 : Termes les plus proches du terme 'younger ones' dans l'EVT

Néanmoins, la cooccurrence de certains mots semble agglomérer certains termes de concept différent. Deux mots apparaissant fréquemment dans des contextes communs se retrouvent alors avec des vecteurs similaires. Cette similarité persiste alors également lors du calcul des vecteurs de termes. C'est par exemple le cas pour les termes relatifs au poisson et ceux relatifs aux fermes de poissons (cf. TABLE 5). Ces représentations sont moins satisfaisantes car elles ne permettent pas de différencier les concepts sous-jacents.

| fish | similarité |
|---|---|
| fish farming | 0,9875 |
| fish farm | 0,9170 |
| disease-free fish farm | 0,9124 |
| fish farm sediments | 0,8683 |
| healthy fish | 0,8145 |

TABLE 5 : Termes les plus proches du terme 'fish' dans l'EVT

## 4.2 Espace vectoriel ontologique (EVO) généré

On peut estimer la qualité des vecteurs de concepts créés en observant la cohérence entre la proximité vecteur/vecteur (en terme de similarité cosinus) et leur position dans l'ontologie.

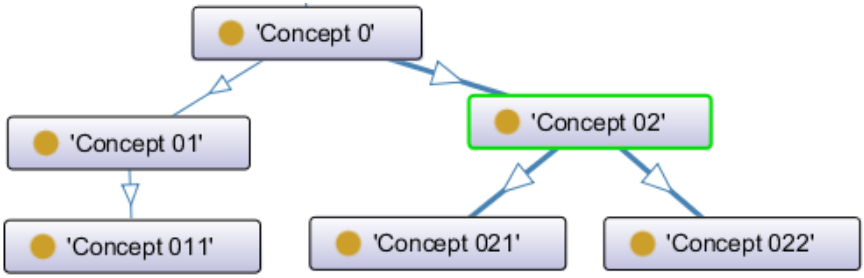



FIGURE 3 : Représentation conceptuelle d'une ontologie (figure générée par le logiciel Protégé (Gennari et al., 2003))

Pour cela, nous pouvons étudier ce qu'il se produit sur une ontologie abstraite (cf. FIGURE 3) :

1. La méthode « One-Hot » : Cette méthode génère des vecteurs de concepts tous orthogonaux entre eux. La similarité cosinus entre 2 vecteurs de concepts distincts est donc égale à 0, qu'il y ait ou non une relation hiérarchique qui les relient (cf. TABLE 6). Cette méthode ne permet donc pas de représenter la structure de l'ontologie.

| **Concept 02** | **similarité** |
|---|---|
| Concept 02 | 1,00 |
| Concept 021 | 0,00 |
| Concept 022 | 0,00 |
| Concept 0 | 0,00 |
| Concept 01 | 0,00 |
| Concept 011 | 0,00 |

TABLE 6 : Proximité du 'Concept 02' aux concepts voisins pour la méthode One-Hot

2. La méthode « Ancestry » : A partir de l'étude des vecteurs de concepts les plus proches d'un vecteur de concept quelconque, il semble ici possible de retrouver les concepts parents et les enfants de celui-ci (cf. TABLE 7). Ensuite, ce sont ses concepts parents puis ses concepts frères qui se retrouvent les plus proches. Il semble alors possible de retrouver l'ensemble de la structure de l'ontologie grâce à l'ensemble des vecteurs de concepts générés par cette méthode.

| **Concept 02** | **similarité** |
|---|---|
| Concept 02 | 1,00 |
| Concept 021 | 0.82 |
| Concept 022 | 0.82 |
| Concept 0 | 0.71 |
| Concept 01 | 0.50 |
| Concept 011 | 0.41 |

TABLE 7 : Proximité du 'Concept 02' aux concepts voisins pour la méthode Ancestry

3. La méthode « Ancestry+ » : Comme attendu, on observe bien un changement dans la proximité des vecteurs de concepts (cf. TABLE 8). Un concept se retrouve plus proche de ses parents que de ses enfants, puis relativement plus éloignés, on retrouve le concept frère plus le concept neveu. On observe également un écartement des concepts entre eux : en terme de

similarité cosinus, les concepts les plus proches se retrouvent à une distance relativement importante vis-à-vis de la méthode précédente.

| Concept 02 | similarité |
|---|---|
| Concept 02 | 1,00 |
| Concept 0 | 0.45 |
| Concept 021 | 0.40 |
| Concept 022 | 0.40 |
| Concept 01 | 0.20 |
| Concept 011 | 0.13 |

TABLE 8 : Proximité du 'Concept 02' aux concepts voisins pour la méthode Ancestry+

4. La méthode « Centered Node » : On retrouve avec cette méthode l'ordre de proximité de la méthode Ancestry (i.e. un concept est d'abord plus proche de ses fils puis de ses parents). Néanmoins, les concepts fils et parents d'un concept courant semblent être à une distance relativement similaire du concept courant. Cette distance semble relativement faible également. Par contre, les concepts frères et neveux se retrouvent beaucoup plus éloignés (cf. TABLE 9).

| Concept 02 | similarité |
|---|---|
| Concept 02 | 1,00 |
| Concept 021 | 0.76 |
| Concept 022 | 0.76 |
| Concept 0 | 0.74 |
| Concept 01 | 0.15 |
| Concept 011 | 0.11 |

TABLE 9 : Proximité du 'Concept 02' aux concepts voisins pour la méthode Centered-Node

## 4.3 Normalisation

Nous allons examiner dans cette section la performance d'un système qui repose sur les espaces vectoriels construits comme expliqué dans les sections précédentes. Rappelons qu'étant donné un terme extrait du texte, sa représentation vectorielle dans l'EVT est obtenue selon l'équation (1), puis son vecteur dans l'EVO est obtenu par la fonction définie en (10) à partir de l'équation (9). Le concept de l'ontologie qui est assigné au terme examiné est alors celui dont le cosinus entre son vecteur et celui de la représentation vectorielle du terme dans l'EVO est maximal.

Pour évaluer la performance des systèmes participants à la tâche de catégorisation de Bacteria Biotope de BioNLP-ST 2016, une mesure de similarité sémantique est implémentée sur le site du challenge BioNLP-ST 2016. La mesure utilisée est celle définie par Wang et al. en 2007 (Wang et al., 2007), avec le paramètre de poids à 0,65. Avec cette mesure, nous pouvons calculer une *baseline* en attribuant à tous les termes le concept « bacteria habitat », qui est la racine de la hiérarchie de l'ontologie OntoBiotope. La mesure trouvée est alors de 32%. Deux équipes ont participé à cette tâche de BioNLP-ST 2016 et ont obtenu les résultats rapportés dans la TABLE 10.

| Equipe | Score |
|---|---|
| BOUN | 0,62 |
| LIMSI | 0,44 |
| Baseline | 0,32 |

TABLE 10 : Performance dans la tâche d'assignation de catégorie de Bacteria Biotope dans BioNLP-ST 2016

On observe une performance globale de nos méthodes au-dessus de la *baseline* (cf. TABLE 10 et TABLE 11). Les 2 meilleures méthodes sont Ancestry et Centered-Node, avec des scores très proches. La moins bonne méthode est Ancestry+, derrière celle qui porte le moins d'information One-Hot.

| Méthode | Score |
|---|---|
| Ancestry | **0,60** |
| Centered-Node | 0,59 |
| One-Hot | 0,55 |
| Ancestry+ | 0,51 |

TABLE 11 : Performance de notre système pour la tâche d'assignation de catégorie de Bacteria Biotope dans BioNLP-ST 2016, en fonction des méthodes de génération d'EVO utilisée

Notre méthode globale avec la méthode de représentation de l'EVO Ancestry a donc obtenu un résultat de 60%, bien au-dessus de la baseline, et très proche de la première équipe (Tiftikci et al., 2016). Ce score est significativement au-dessus de la méthode du LIMSI (Grouin, 2016), méthode basée sur une approche morphologique.

# 5 Discussions

La qualité globale des espaces vectoriels générés n'a été évaluée que via la tâche de normalisation présentée. Il serait pertinent d'examiner les effets du système avec d'autres tâches. Nous prévoyons d'effectuer cela dans de futurs travaux.

Les résultats inférieures à la méthode One-Hot de la méthode Ancestry+ peut laisser supposer que la réorganisation de l'espace opéré par cette méthode est soit non-pertinente, soit trop brutale (cf. répartition en distance des concepts dans la TABLE 8). A l'inverse, la One-Hot reste une méthode aux résultats honorables alors qu'elle n'apporte aucune information sur la structure de l'ontologie. Ce résultat montre qu'avec l'ajout de l'information sur la structure de l'ontologie, les représentations Ancestry et Centered-Nodeont ont pu apporter un gain de 4-5 points dans la tâche. Reste à savoir s'il existe des représentations qui permettraient d'améliorer encore plus ces résultats, ou si l'on se rapproche d'un certain palier.

Un des plafonds de la méthode présentée dans cet article est dû au fait que, pour cette tâche de normalisation, un terme peut être normalisé par plusieurs concepts de l'ontologie (ex : le terme « school age children with wheezing illness » devrait être normalisé par le concept <OBT:002307: pediatric patient> ainsi que le concept <OBT:002187: patient with disease>).

# 6 Conclusion et perspectives

L'objectif de cet article était de proposer une approche pertinente et originale pour la création de représentations vectorielles pour des termes (complexes ou non) dans un espace sémantique pertinent pour une tâche d'assignation de concepts à des termes désignant des entités dans un corpus. De plus, il visait à proposer une méthode capable de s'adapter à un corpus spécialisé de petite taille où les termes d'intérêt ont un faible nombre d'occurrences. Les résultats, proches des meilleurs des participants à la tâche de catégorisation de BioNLP-ST 2016, semble ouvrir d'encourageantes perspectives.

Pour de futurs travaux, il serait pertinent d'appliquer des méthodes d'évaluation globale de la qualité des espaces vectoriels générés. En particulier, cela permettrait d'évaluer plus exhaustivement les processus intermédiaires et d'observer avec plus de précision l'impact des modifications sur leurs paramètres internes. De nouvelles méthodes plus élaborées pourraient alors être envisagées pour améliorer les résultats. Par exemple, la méthode utilisée ici pour générer les vecteurs de l'EVT pourrait être améliorée pour prendre en compte la structure syntaxique des termes. Cela pourrait résoudre les problèmes de similarité sémantique entre « fish » et « fish farm » (cf. TABLE 5). Plus généralement, il serait intéressant de comparer les effets d'autres méthodes de génération de vecteurs de termes sur la performance du système présenté.

Appliquer cette même méthode avec une ontologie dont tous les concepts sont définis pourrait être bénéfique. Si de telles ontologies sont rares, il pourrait être tout de même intéressant de tester en ne prenant en compte qu'une partie des concepts dont les intersections avec d'autres ne sont pas définies.

Enfin, malgré la limitation inhérente des méthodes de normalisation basées sur la morphologie des mots, celles-ci pourraient néanmoins être utilisées pour effectuer une pré-normalisation du corpus. En conséquence, on pourrait envisager d'utiliser ces annotations pour entraîner la méthode au lieu d'utiliser une annotation manuelle. Ainsi, cela transformerait cette méthode en une méthode non-supervisée.

# Remerciements

This work is supported by the "IDI 2015" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

# Références

DELGER L., CHAIX E., BA M., FERRE A., BESSIERES P., NEDELLEC C. (2016). Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016.

EIDOON Z., YAZDANI N., OROUMCHIAN F. (2007). A Vector Based Method of Ontology Matching (p. 378‑381). IEEE.

FABRE C., HATHOUT N., SAJOUS F., TANGUY L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)* (p. 266–279).

FABRE C., LENCI A. (2015). Distributional Semantics Today Introduction to the special issue. *Traitement Automatique des Langues*, *56*(2), 7–20.

FIRTH J. R. (1957). The technique of semantics. *Oxford University Press*. London.

GENNARI J. H., MUSEN M. A., FERGERSON R. W., GROSSO W. E., CRUBÉZY M., ERIKSSON H., … TU S. W. (2003). The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, *58*(1), 89–123.

GOLIK W., WARNIER P., NÉDELLEC C. (2011). Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence* (p. 37–39).

GROUIN C. (2016). Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts. *ACL 2016*, 64.

HARRIS Z. S. (1954). Distributional Structure. WORD, *10*(2‑3), 146‑162.

MANIEZ F. (2007). Prémodification et coordination : quelques problèmes de traduction des groupes nominaux complexes en anglais médical. *ASp*, (51‑52), 71‑94.

MCCRAY A. T. (1989). The UMLS Semantic Network. In *Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care* (p. 503–507). American Medical Informatics Association.

MIKOLOV T., CHEN K., CORRADO G., DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

NAZARENKO A., NEDELLEC C., ALPHONSE E., AUBIN S., HAMON T., MANINE A.-P. (2006). Semantic annotation in the alvis project. In *International Workshop on Intelligent Information Access (IIIA)* (p. 5–pages).

TIFTIKCI M., SAHIN H., BÜYÜKÖZ B., YAYIKÇI A., OZGÜR A. (2016). Ontology-based Categorization of Bacteria and Habitat Entities using Information Retrieval Techniques. *ACL 2016*, 56.

WANG J. Z., DU Z., PAYATTAKOOL R., YU P. S., CHEN C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274‑1281.

# Annotation automatique des lieux dans l'oral spontané transcrit

Hélène Flamein[1]
(1) LLL, UMR 7270, UFR LLSH, 10 Rue de Tours, 45065 ORLEANS cedex 2
helene.flamein@univ-orleans.fr

## RESUME

Cet article a pour but de présenter une démarche généraliste pour l'annotation automatique des lieux dans l'oral transcrit. Cette annotation est effectuée sur le corpus ESLO (Enquête SocioLinguistique à Orléans) et suppose une réflexion sur les caractéristiques propres à la désignation d'un lieu à l'oral. Avant d'expliciter la méthode employée pour traiter automatiquement notre corpus, nous présenterons le travail préparatoire de la constitution d'une convention d'annotation et d'un corpus de référence indispensable pour l'évaluation du système.

## ABSTRACT

**Automatic annotation of places in the transcribed oral**

The aim of this article is to present a general methodology for the automatic annotation of places in the transcribed oral. This annotation is carried out on the corpus ESLO (Enquête SocioLinguistique à Orléans – Sociolinguistic Investigation in Orléans) and presupposes a reflection on the characteristics proper to the designation of an oral place. Before explaining the method used to automatically process our corpus, we will present the preparatory work for the constitution of an annotation convention and a corpus of reference indispensable for the evaluation of the system.

MOTS-CLES : Détection automatique de lieux, Perception des lieux, Traitement automatique du Langage, Entités nommées, ESLO, Corpus oral

KEYWORDS: Automatic detection of places, Place-perception, Natural Language Processing, Named entities, ESLO, Oral corpus

## 1 Introduction

Cette étude s'inscrit dans un travail de thèse qui s'intéresse à la perception des habitants d'une ville. L'objectif à terme est de présenter l'analyse automatique de la perception des lieux dans le discours transcrit de l'oral par différents locuteurs grâce aux techniques du TAL. Ces techniques permettent notamment de détecter automatiquement les lieux ainsi que les expressions qui les accompagnent dans les transcriptions. Nous appuyons notre travail sur le corpus ESLO2 (Enquête SocioLinguistique à Orléans[1]), composé de 460 heures d'enregistrements répartis en une vingtaine de modules qui présentent chacun des situations de communication différentes.

[1] http://eslo.huma-num.fr/

Après avoir fait le rappel du cadre théorique encadrant notre démarche, nous présenterons la méthode suivie pour le repérage des lieux. Ses différentes étapes seront développées : l'exploration manuelle du corpus, l'élaboration de conventions d'annotation, la constitution d'un corpus de référence ainsi que notre expérience d'annotation automatique. A la fin du processus d'annotation, les lieux détectés seront projetés sur la carte d'Orléans grâce à leurs coordonnées géographiques présentes dans les métadonnées des ressources spécifiques constituées.

# 2 Etat de l'art

L'utilisation de corpus annotés est une pratique fondamentale en TAL et ce depuis les années 90. Le corpus arboré de l'anglais de l'Université de Pennsylvanie créé entre 1989 et 1994 ou Penn Treebank fait référence avec ses annotations morpho-syntaxiques et syntaxiques en inspirant la création d'autres ressources similaires telles que le corpus arboré du français présenté par Abeillé et al. (2003).

Depuis les années 1990 et la dernière série des conférences américaines MUC[2] (Message Understanding Conferences), la question de la reconnaissance des entités nommées est incontournable dans le domaine du TAL. Selon (Ehrmann, 2008), les entités nommées correspondent à « toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Elles représentent des objets textuels porteurs de sens généralement classés selon plusieurs catégories : lieux, personnes, organisations, dates, unités monétaires et pourcentages (Maurel et al., 2011 ; Nadeau et Sekine, 2009). De nombreux outils d'annotation automatique sont dédiés à la reconnaissance des entités nommées (REN) comme l'un des modules du Stanford NER[3] (Finkel et. al, 2005) ou le POLYGLOT NER[4] présenté dans Al-Rafou et. al (2015). La REN est devenue une tâche indépendante qui est désormais au centre de différentes campagnes d'évaluation d'outils dédiés à l'extraction d'informations. Plusieurs campagnes comme ESTER ou ETAPE évaluent justement l'annotation des entités nommées dans des corpus d'émissions radiophoniques ou télévisuelles. Ces projets présentent chacun des conventions d'annotation des entités nommées qui ont inspiré notre propre convention d'annotation des lieux.

L'objectif du travail est d'annoter l'ensemble des mentions de lieux présentes dans l'oral transcrit. Les lieux correspondent ici à la définition de Lesbeguerrie (2007) des entités spatiales absolues (ESA) qui représentent les informations spatiales les plus « primitives » et les plus proches de la définition des entités nommées de type lieux (ex : la ville d'Orléans, le campus de la Source). Les ESA associées à des indications géographiques (ex : au sud de la ville d'Orléans, près du campus de la Source) sont qualifiés d'entités spatiales relatives (ESR). A la différence des projets de REN mentionnés précédemment, notre analyse se fonde sur un corpus oral spontané. Comme démontré par Brando et. al, (2016), ce type de corpus peut être riche en entités nommées mais surtout en mentions de lieux génériques (ex : un endroit très beau, le long de la grande rue). Pour reconnaître ces types de lieux qui se rapprochent des ESR, les systèmes de REN existants ne sont pas adaptés au contexte oral. Dans l'optique de faire l'analyse de la perception des entités spatiales identifiées dans notre corpus, nous privilégions la constitution d'un nouveau système de reconnaissance des lieux qui préparera directement l'analyse de la perception de ces mêmes lieux.

[2] http://www-nlpir.nist.gov/related_projects/muc/
[3] http://nlp.stanford.edu/software/CRF-NER.html
[4] http://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html

# 3 Présentation des données exploitées

## 3.1 Le projet ESLO

Le corpus traité est le corpus ESLO (Eshkol-Taravella et al. 2012). Ce corpus comprend différentes situations d'enregistrement (entretiens face à face, interviews de personnalités, enregistrements dans des cours de récréation, pendant des repas, etc.).

Chaque enregistrement est transcrit orthographiquement avec une distinction entre les tours de parole sans signe de ponctuation, les points d'interrogation et les majuscules pour les noms propres étant les seules exceptions admises.

## 3.2 Modules exploités

Deux modules du corpus ESLO2 ont été sélectionnés : Entretiens et Itinéraires.

Lors de l'entretien, le locuteur, un habitant d'Orléans ou de son agglomération, est amené à faire état de son histoire personnelle, à partager ses habitudes de vie, etc. Les locuteurs témoins expriment ainsi à travers les entretiens leur perception de la ville d'Orléans. 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. 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 très 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.

Ces deux modules constituent une ressource riche en mentions de lieux relatives à Orléans. Bien qu'elle servira à entraîner notre système de détection automatique des mentions de lieux, notre méthode reste généraliste dans le sens où elle doit être applicable sur l'ensemble du corpus ESLO.

# 4 Méthode

Dans l'objectif de repérer les lieux présents dans l'oral transcrit, notre démarche (cf. Figure 1) comprend 4 étapes suivantes : (1) Constitution d'un corpus d'entraînement pour l'exploration manuelle des données, (2) Création de convention d'annotations, (3) Constitution d'un corpus de référence et (4) Expérience d'annotation automatique des lieux.

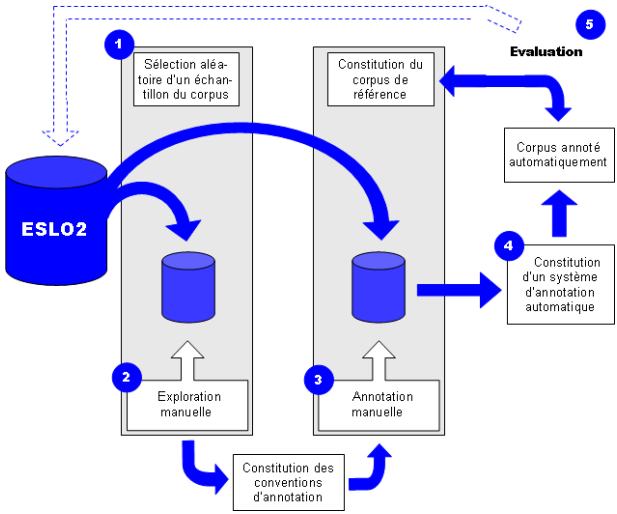



FIGURE 1 : Démarche globale pour le repérage des lieux

## 4.1 Exploration manuelle du corpus

Une dizaine de transcriptions issues des modules Entretiens et Itinéraires ont été sélectionnées pour l'examen manuel des types de noms de lieux présents dans le corpus ce qui a permis d'établir une première typologie des lieux (Eshkol-Taravella et Flamein, à paraître) et de constituer des conventions d'annotation en adéquation avec notre corpus.

Les noms de lieux peuvent eux-mêmes être désignés de manières très variées et parfois personnelles. Ils peuvent être abrégés, tronqués, transformés par le locuteur. Un même lieu peut aussi être mentionné sous différents noms sans que cela n'en gêne le sens comme dans :

1. *ce bout de d'île euh qui est euh pas très loin du pont euh euh du **pont George Cinq** oui c'est **le pont Royal*** (ESLO2_ENT_1034_C)

*Pont George Cinq* est le nom officiel actuel de l'un des ponts d'Orléans. Il a été inauguré en 1763 sous le nom de *Pont Royal*. Même s'il a changé plusieurs fois de noms entre temps, il est très régulier que les Orléanais se réfèrent à celui-ci en utilisant son ancien nom. L'enjeu de notre travail est de détecter toutes les mentions de lieux présentes dans le corpus tout en prenant en compte la variation dans leur dénomination. Le système de l'annotation automatique doit être capable de lier une entité nommée, nom officiel du lieu, à ses possibles variantes pour permettre 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. La prise en compte de ses éléments se traduit directement dans la convention d'annotation présentée dans la partie suivante.

## 4.2 Conventions d'annotation des mentions de lieux

Pour annoter l'ensemble des mentions de lieux présentes dans le corpus, nous utilisons la balise XML : <loc> … </loc> à laquelle nous ajoutons trois attributs. Ces attributs préciseront la nature des lieux identifiés et serviront de guide à l'analyse future de la perception.

### *4.2.1 Typologie des lieux*

La convention d'annotation proposée s'inspire de celle des campagnes ESTER2[5] et ETAPE[6] (Rosset, Grouin et Zweigenbaum, 2011). ETAPE a pour objectif d'évaluer les performances des technologies vocales appliquées à l'analyse de flux télévisés en langue française tandis qu'ESTER2 est un projet antérieur à ETAPE avec des objectifs similaires de mesure de performances de systèmes de transcriptions d'émissions radiophoniques.

En vue de l'analyse de l'image d'une ville, il nous est nécessaire de conserver une typologie fine des lieux urbains. Une finesse que nous ne conservons pas pour la description des lieux naturels, extra-urbain. Notre convention reprend les codes de la convention Quaero en ce qui concerne les lieux que l'on peut découper administrativement (quartiers, villes, pays…). Les lieux géographiques seront considérés dans leur ensemble comme dans la typologie de la convention ESTER2.

| **<loc type="  ">** | |
|---|---|
| Villes | type="ville" |
| *Orléans, Paris, La Ferté-St-Aubain, Dunois, La Source…* | |
| Région | type="region" |
| *Loiret, région Centre Val-de-Loire, Beauce, Gâtinais…* | |
| Pays | type="pays" |
| *France, Espagne, Royaume-Uni, Chine…* | |
| Supranational | type="supra" |
| *Europe, Asie du Sud Est, le Nord, la Flandre…* | |
| Rues, avenues, ponts… | type ="voie" |
| *rue de la République, Pont Royal…* | |
| Lieux physiques naturels | type ="naturel" |
| *Forêt d'Orléans, Loire, Canal de Briare…* | |

TABLE 3 : Nouvelle typologie des lieux (1/2)

Les lieux considérés comme des constructions humaines et les organisations sont des entités souvent difficiles à distinguer. Le contexte d'emploi influence la catégorie dans laquelle classer l'entité observée. Ainsi dans les exemples [2] et [3] :

[5] http://www.afcp-parole.org/camp_eval_systemes_transcription/
[6] http://www.afcp-parole.org/etape.html

2. *alors euh je suis étudiante à l'**université euh d'Orléans La Source***

3. *comme je vais à la à l'**université** à La Source ça me fait quand même de la route* (ESLO2_ENTJEUN_03_C)

Le même locuteur mentionne à deux reprises *l'université d'Orléans* et le quartier *La Source*. Dans l'exemple [2], l'université est clairement désignée en tant qu'organisation alors que dans l'exemple [3], il s'agit d'un lieu. Que ce soit dans les conventions proposées par ESTER2 ou Quaero, l'*université d'Orléans* serait catégorisée comme une construction humaine dans l'exemple [3]. Aucune distinction n'est prévue au niveau du type de construction. L'université, la mairie ou toute autre structure commerciale sont catégorisées de la même façon. Pour une analyse plus fine de la perception des lieux de la ville d'Orléans, nous souhaitons préciser la nature des constructions humaines identifiées. Pour ce faire, nous nous référons à la typologie des entités nommées de type organisation définie dans les conventions ESTER2 et Quaero.

| **Organisations** | |
|---|---|
| Politique | org.pol |
| Educative | org.edu |
| Commerciale | org.com |
| Non commerciale | org.non-profit |
| Média & divertissement | org.div |
| Géo-socio-administrative | org.gsp |

TABLE 5 : Typologie des entités nommées de type organisation selon ESTER2

A partir du contexte d'emploi, le système développé devra déterminer s'il a affaire à un lieu ou à une organisation. Il pourra s'appuyer sur la présence de verbes de déplacements ou sur l'association de certaines prépositions comme *à* ou *vers* et de certains verbes intrinsèquement liés à la notion de lieux : *habiter, naître, vivre*… Dans les cas où il détectera une organisation, aucune annotation ne sera faite. Si une construction humaine est identifiée, celle-ci sera annotée et considérée comme un lieu auquel on attribut une fonction. La typologie des fonctions possibles des lieux reprend celle des organisations dans la convention ESTER2 (cf. table [5]).

| **<loc type=" ">** | |
|---|---|
| Lieux à dimension historique, touristique | type ="monument" |
| *Cathédrale Sainte Croix, Hôtel Groslot…* | |
| Lieux à fonction administrative | type ="admin" |
| *Mairie d'Orléans, Office du Tourisme, CAF…* | |
| Lieux à fonction éducative | type ="educatif" |
| *Lycée Pothier, Université d'Orléans…* | |
| Lieux à fonction commerciale | type ="commerce" |
| *Carrefour, H&M, Menphis Coffee…* | |
| Lieux à fonction non commerciale | type ="ncommerce" |
| *Hôpital de la Source, Secours Populaire,…* | |

TABLE 6 : Nouvelles typologie des lieux (2/2)

Notre typologie conserve les catégories principales proposées par ESTER2 et les constructions humaines sont typées de façon similaire aux organisations (cf. table [6]). 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, ministères, etc.) ou les organisations militaires reliées au gouvernement (ex : CIA, Marine Nationale…), etc. Nous ne conservons pas ce type puisque nous considérons que les organisations avec une fonction politique. Si des lieux à fonction politique sont évoqués, ils seront plutôt inclus avec le type « admin » de notre convention.

En conclusion, si l'on reprend l'exemple [3], le quartier de la Source et l'université d'Orléans seront annotés selon notre convention de la façon suivante :

3. *comme je vais à la à l'*<loc type="educatif" zone="2" label="université d'Orléans">***université***</loc> *à* <loc type="ville" zone="2" label="Orléans-la-Source">*La Source*</loc> *ça me fait quand même de la route* (ESLO2_ENTJEUN_03_C)

### *4.2.2 Zone géographique*

Trois zones géographiques sont distinguées dans l'annotation. Elles correspondent 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…* | |

TABLE 7 : Zone géographique

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.

4. *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)

### *4.2.3 Label officiel*

L'attribut du label officiel correspond au nom officiel du lieu identifié. Cet attribut répond à la capacité à varier du nom du lieu. Lorsqu'un locuteur mentionne un lieu, il peut se l'approprier, le personnifier en opérant des modifications sur son nom (troncations, utilisation de surnom…). Le nom officiel du lieu correspond à sa forme complète, sans aucune modification du locuteur.

5. *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)*

6. *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)*

Dans les exemples [3] et [4] les lieux *rue de la République* et *La Ferté-Saint-Aubin* ont été tronqués par le locuteur (*rue de la Rép-* et *La Ferté*). L'attribut label prend pour valeur la forme complète du nom de ces lieux.

L'intérêt de faire figurer le nom officiel dans la balise du lieu réside dans notre volonté de géoréférencer les lieux identifiés. Cette information servira à rechercher dans une base de données les coordonnées géographiques du lieu pour le placer sur la carte finale.

## 4.3 Annotation manuelle

Les conventions d'annotation définies, nous procédons à l'annotation maneulle du corpus pour constituer un premier corpus de référence. Le corpus de référence est constitué de 4 transcriptions aléatoirement sélectionnées parmi les 91 transcriptions du module Itinéraires différentes de celles analysées dans la première étape d'observation. Ce module a été privilégié puisque plus concentré en mentions de lieux proportionnellement à sa taille par rapport aux transcriptions du module Entretien. Cette concentration offre un gain de temps et d'efficacité au moment de l'annotation manuelle.

| Transcriptions | Durée | Nombre de mots |
|---|---|---|
| ESLO_iti_02_01 | 02:39 | 570 |
| ESLO_iti_02_03 | 03:20 | 704 |
| ESLO_iti_02_05 | 04 :01 | 893 |
| ESLO_iti_02_06 | 12:23 | 2815 |
| **Totaux** | 22:23 | 4982 |

TABLE 8 : Volume de données par transcriptions du corpus de référence

Afin de tester l'efficacité de la convention d'annotation, nous avons procédé à une évaluation de l'accord inter-annotateur avec la mesure du Kappa de Cohen. Cette évaluation concerne en particulier le calcul d'accord sur le type de lieu choisi par les annotateurs lorsqu'ils ont détecté le même lieu. Dans le cadre d'un cours de master, notre propre version annotée (A1) du corpus présenté en table [8] a été comparé à celle d'une étudiante (A2). Le Kappa de Cohen permettra de montrer l'accord des deux juges en ce qui concerne le choix des types de lieux.

Pour procéder au calcul, nous utilisons l'application mise à disposition librement en ligne par Philippe Bonnardel[7]. En suivant la procédure, nous obtenons un Kappa égal à 0.810. Si l'on se rapporte à la grille de lecture proposée par (Landis et Koch, 1977), l'accord est considéré comme

[7] http://kappa.chez-alice.fr/

excellent. Cette première évaluation montre un accord élevé entre les annotateurs mais laisse paraître quelques désaccords. L'évaluation de l'annotation s'est ensuivie d'une phase d'adjudication durant laquelle chaque désaccord a été analysé. Par exemple, dans :

7. A1 : *la mairie la* **<loc type="admin" zone="2" label="mairie d'Orléans">*mairie d'Orléans*</loc>** *elle est belle*

   A2 : *la mairie la* **<loc type="monument" zone="2" label="mairie d'Orléans ">*mairie d'Orléans*</loc>** *elle est belle*

A1 et A2 sont en désaccord sur la façon de catégoriser le lieu *mairie d'Orléans*. Même si le bâtiment abritant la mairie peut avoir un intérêt touristique, nous considérons que la fonction administrative prévaut sur l'attrait touristique de la structure. Le type "admin" sera donc privilégié.

Après comparaison des deux annotations et résolution des désaccords, nous avons pu aboutir à une version consensuelle. Cette troisième version annotée constitue notre corpus de référence. Ce corpus annoté sera utile pour l'évaluation de notre système.

## 4.4 Expérience d'annotation automatique

### *4.4.1 Méthodologie*

Pour repérer des lieux, nous utilisons les méthodes heuristiques prenant en compte leur contexte d'emploi dans le discours. Le système d'annotation est en cours d'élaboration.

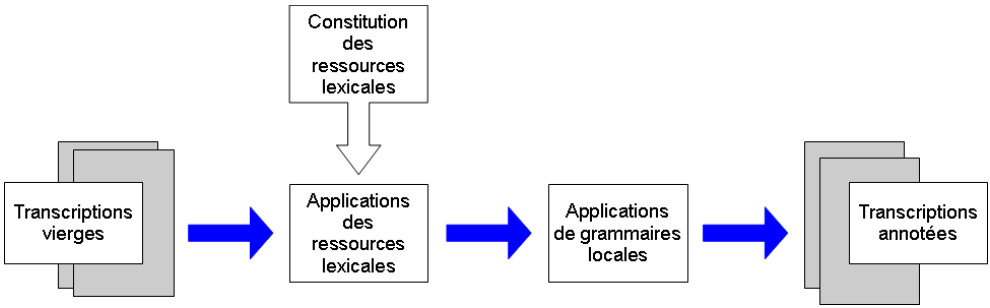



FIGURE 2 : Chaîne de traitement automatique du corpus

L'annotation des lieux dans le corpus repose dans un premier temps sur l'utilisation de ressources lexicales dédiées à l'information géographique. En complément de ces ressources, différentes règles sont appliquées afin d'identifier toutes les mentions de lieux absentes des listes.

#### *4.4.1.1 Ressources utilisées*

Les ressources utilisées sont constituées à partir de bases de données disponibles en ligne et dédiées à l'information géographique comme GEOFLA®[8] qui contient la liste de tous les noms de voies de circulation, de communes, de départements, et de régions en France. Ces bases de données géographiques permettront l'identification de tous les lieux mentionnés par leur nom officiel, sans modification apportées par le locuteur. De plus, elles pourront renseigner l'attribut de la zone géographique à renseigner dans l'annotation.. Ces bases de données mettent aussi à disposition des coordonnées géographiques de chaque lieu répertorié. Grâce à cette information, nous serons capables de placer sur une carte tous les lieux identifiés dans les transcriptions analysées.

#### *4.4.1.2 Variations des noms de lieux*

Le nom d'un lieu peut varier et peut être mentionné sous un nom complètement différent du son nom d'origine.

Pour gérer ces variations dans la désignation des lieux, nous utilisons notamment la distance de Levenshtein, dans la lignée de Zesnani et al. (2005), qui l'avaient utilisée pour l'enrichissement d'un dictionnaire d'entités spatiales. Cette distance mathématique permet de mesurer la similarité entre deux chaînes de caractères et est exploité dans le projet pour gérer les cas d'abréviation des noms de lieux comme dans les exemples (3) et (4). C'est la distance de Levenshtein qui doit permettre d'identifier *La Ferté* comme la version abrégée de *La Ferté-Saint-Aubin*, nom renseigné dans la base.

Cette méthode connaît toutefois des limites. Pour décider si les deux chaînes comparées renvoient à la même entité, il faut définir un seuil afin de déterminer si la distance obtenue est trop grande ou non. Ce seuil est généralement défini en fonction de la longueur de la chaîne observée. En l'état, la distance amène autant de bonnes détections que de bruits dans l'annotation. Si l'on reprend l'exemple de la chaîne de caractères *La Ferté* comparée à la chaîne *La Ferté-Saint-Aubin*, la distance de Levenshtein vaut 12. La différence entre les deux chaînes est supérieure à la taille de la chaîne *La Ferté* (8 caractères). Le seuil définit dans le système demandera automatiquement le rejet de *La Ferté* comme une variante de *La Ferté-Saint-Aubin* alors que celle-ci est pertinente.

Pour contrer cette difficulté, il est nécessaire d'ajouter d'autres règles pour compléter l'analyse. L'une des premières pistes pour pallier à ce problème est l'adaptation même de la distance pour qu'elle ne raisonne plus en nombre de caractères différents mais plutôt en nombre de mots. La combinaison de la distance de Levenshtein classique et de sa version adaptée pourrait permettre d'améliorer la qualité de l'annotation. Dans le cas de *La Ferté* et *La Ferté-Saint-Aubin*, la distance en mots serait de 2 et l'utilisation classique de la distance pourra déterminer que deux mots sont identiques dans les deux chaînes. Ces deux informations pourront permettre de valider *La Ferté* comme une variante de *La Ferté-Saint-Aubin*.

Cependant, le seuil à définir doit rester suffisamment strict afin de limiter, voir éliminer le bruit. Les conclusions de la distance devront être confirmés ou infirmés par d'autres règles basées sur le contexte ou sur la nature des mots constituant l'entité observée.

[8] Base de données mise à disposition librement par l'Institut National de l'Information Géographique et Forestière (IGN) - http://professionnels.ign.fr/geofla

#### *4.4.1.3 Evaluation du système*

L'évaluation du système n'a pas encore été réalisée puisque le système d'annotation est encore en cours d'élaboration. Si l'annotation des noms de lieux simples (c'est-à-dire les entités nommées sans variations) sont déjà identifiées grâce aux ressources lexicales, beaucoup d'autres mentions comme les noms de lieux génériques ou les noms soumis à variations ne le sont pas encore. Il est nécessaire de continuer le développement de l'outil pour proposer une évaluation pertinente de celui-ci.

# 5 Perspectives

## 5.1 Poursuite de l'élaboration du système

Comme évoqué précédemment, il est nécessaire d'ajouter d'autres règles dans le système afin d'améliorer notamment l'efficacité de la distance de Levenshtein mais surtout pour permettre l'identification des les lieux absents des bases de données utilisées. Ce problème concerne en particulier les lieux avec une activité commerciale ou les lieux publics pour lesquels nous ne disposons pas de listes similaires aux bases de GEOFLA®. Ces lieux font partie intégrante du quotidien des locuteurs interrogés. On peut alors supposer que les locuteurs vont plus naturellement agir sur le nom du lieu auquel ils font référence. Les phénomènes les plus courants sont l'abréviation ou l'utilisation de variantes du nom officiel. L'enjeu est ici de reconnaître le lieu sans distinction du procédé utilisé pour le mentionner. La difficulté principale dans le repérage est de tenir compte de toutes les variations possibles et de tenir compte de la coréférence entre plusieurs mentions du même lieu. Il sera nécessaire de constituer des lexiques dédiés à l'information géographique en listant par exemple tous les noms communs représentants des lieux. Ces listes ou dictionnaires pourront être combinés à des règles syntaxiques prenant en compte le contexte d'emploi des lieux.

Il faudra aussi être capable de résoudre les cas d'homonymie provoquée par l'abréviation d'un nom de lieu. Par exemple, l'agglomération orléanaise englobe notamment les villes Saint-Jean-de-la-Ruelle, Saint-Jean-de-Braye et Saint-Jean-le-Blanc. commeDans l'exemple :

8. *EW15: #1 oui c'était un à **Saint Jean** #*
   *ch_AC7: #2 c'était où à ? # à **Saint Jean-le-Blanc** ?*
   *EW15: **Saint Jean-le-Blanc** #1 oui #*
   *ch_AC7: #2 ah oui*
   (ESLO2_ENT_1015_C)

à laquelle de ces trois villes renvoie la mention *Saint Jean* ? Il sera nécessaire de prendre en compte le contexte d'énonciation et les tours de parole précédents afin de désambigüiser ce type de cas. On peut supposer qu'avant d'utiliser la forme abrégée du nom de la ville, le locuteur aura utilisé la forme complète afin de s'assurer que son interlocuteur le comprenne. En interrogeant le reste de la transcription et en associant la recherche à la distance de Levenshtein, le système pourrait identifier le nom de ville le plus proche en termes de position dans la transcription et d'orthographe. Dans l'exemple [7], *Saint Jean-le-Blanc* serait alors identifié comme la forme complète de *Saint Jean*.

### 5.2 Module de géoréférencement

L'objectif final du travail est de représenter les lieux et les avis qui les concernent sur une carte. Dans cette optique, nous allons intégrer à notre système un module de géoréférencement. Nous pourrons nous appuyer sur les ressources lexicales constituées et les bases de données dédiées à l'information géographique dans lesquelles sont renseignées ces coordonnées. Les coordonnées seront incluses dans un Système d'Information Géographique et chacun des lieux d'Orléans ou de son agglomération pourront être placés les uns par rapport aux autres sur la carte. Il ne sera pas pertinent de représenter les lieux en dehors de ces zones, d'où la nécessité d'obtenir cette information au moment du repérage.

### 5.3 Analyse de la perception des lieux

Le repérage des lieux est la première étape de notre travail. A partir de ce repérage, nous serons en mesure de faire l'analyse de la perception de ces lieux. La dénomination d'un lieu est un processus social réapproprié subjectivement et est déterminée par la personnalité, l'histoire du locuteur. Un lieu peut être apprécié, ou non, par le locuteur. Au niveau de l'analyse des variations dans les noms de lieux, nous pouvons nous rapprocher de la notion des lieux subjectifs développée par Dominguès et Eshkol (2013, 2015). Mentionner un lieu n'est pas un processus neutre et les moyens pour exprimer ses sensations à son sujet sont multiples comme cela a été démontré dans Eshkol-Taravella et Flamein (à paraître).

## 6 Conclusion

Nous proposons une démarche pour la création d'un module de reconnaissance automatique des lieux qui se veut généraliste puisqu'il est applicable sur toutes les transcriptions du corpus ESLO. Les perspectives du travail présenté est d'offrir une visualisation cartographique de la perception des Orléanais de leur lieu de vie. Pour ce faire, nous avons défini une convention d'annotation ce qui nous a permis de constituer un corpus de référence dédié à l'évaluation de notre système d'annotation automatique.

A terme, les expressions subjectives identifiées dans les transcriptions analysées formeront la vision que des Orléanais ont de leur propre ville. Cette vision sera comparée avec celle proposée par les données issues des bases de données Linked Open Data comme Wikipédia. Le fait de prendre en compte les témoignages d'ESLO et d'y associer l'ensemble des informations neutres et objectives des bases de données donnera une dimension anthropologique et sociologique à la carte produite et permettra de constituer un véritable portrait de la ville d'Orléans.

# Références

ABEILLE A., CLEMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Anne Abeillé*, éditeur : Treebanks, pages 165 –187. Kluwer, Dordrecht.

BRANDO C., DOMINGUES C., CAPEYRON M. (2016). Evaluation of NER systems for the recognition of place mentions in French thematic corpora, In: *Proceedings of the 10th Workshop on Geographic Information Retrieval (GIR '16)*. ACM, New York, NY, USA, article 7, 10 p ages DOI: 10.1145/3003464.3003471

COHEN J. (1960). A coefficient of agreement for nominal scales., *Educ. Psychol. Meas*., 20, 27-46.

DOMINGUES C., ESHKOL-TARAVELLA I. (2015). Toponym recognition in custom-made map titles. *International Journal of Cartography*, Volume 1, Taylor & Francis.

DOMINGUES C., ESHKOL-TARAVELLA I. (2013). Repérer des toponymes dans les titres de cartes topographiques. *TALN2013*, Jun 2013, Les Sables d'Olonne, France. 2013,

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).

ESHKOL-TARAVELLA I., FLAMEIN H. (à paraître)

FORT K. (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.

KERGOSIEN E., MAUREL P., ROCHE M., TEISSEIRE M. (2015). Senterritoire pour la détection d'opinions liées à l'aménagement d'un territoire. *Revue Internationale de Géomatique*, Hermes, 25 (1), pp.11-34.

LANDIS J.R., KOCH G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, (33) :159–174, 1977.

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.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I. & NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *TAL*, (vol. 52, n° 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).

FINKEL J. R., GRENAGER T. & MANNING C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

AL-RAFOU R., KULKARNI V., PEROZZI B. & SKIENA S. (2015). Polyglot-NER : Massive Multilingual Named Entity Recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining*, Vancouver, British Columbia, Canada, April 30 - May 2, 2015.

ROSSET S., GROUIN C., ZWEIGENBAUM P. (2011). *Entités Nommées Structurées : guide d'annotation Quaero*. Technical report.

ZENASNI S., KERGOSIEN E., ROCHE M, TESSEIRE M. (2016). Extracting new Spatial Entities and Relations from Short Messages, In the *8th International ACM Conference on Management of Digital EcoSystems (MEDES'2015)*, pp. 8, Hendaye (France).

# Vers la détection automatique des affirmations inappropriées dans les articles scientifiques

Anna Koroleva[1]
(1) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
koroleva@limsi.fr

## RESUME

Dans cet article nous considérons l'apport du Traitement Automatique des Langues (TAL) au problème de la détection automatique de « l'embellissement » (en anglais « spin ») des résultats de recherche dans les publications scientifiques du domaine biomédical. Nous cherchons à identifier les affirmations inappropriées dans les articles, c'est-à-dire les affirmations où l'effet positif du traitement étudié est plus grand que celui effectivement prouvé par la recherche. Après une description du problème de point de vue du TAL, nous présentons les pistes de recherche qui nous semblent les plus prometteuses pour automatiser la détection de l'embellissement. Ensuite nous analysons l'état de l'art sur les tâches comparables et présentons les premiers résultats obtenus dans notre projet avec des méthodes de base (grammaires locales) pour la tâche de l'extraction des entités spécifiques à notre objectif.

## ABSTRACT

**Towards automatic detection of inadequate claims in scientific articles**
In this article we consider application of Natural Language Processing (NLP) techniques to the task of automatic detection of misrepresentation (« spin ») of research results in scientific publications from the biomedical domain. Our objective is to identify inadequate claims in medical articles, i.e. claims that state the beneficial effect of the experimental treatment to be greater than it is actually proven by the research results. After analyzing the problem from the point of view of NLP, we present methods that we consider applicable for automatic spin identification. We analyze the state of the art in similar or related tasks and we present our first results obtained with basic methods (local grammars) for the task of recognising entities specific for our goal.

MOTS-CLES : affirmations inappropriées ; spin ; embellissement de résultat, articles biomédicaux ; rôles sémantiques ; extraction des entités

KEYWORDS : inadequate reporting ; spin ; biomedical articles ; semantic roles ; entity extraction

# 1 Introduction

L'interprétation des résultats de recherche scientifique dans le domaine biomédical est souvent affectée par la présence d'embellissements ou d'affirmations inappropriées. Embellir les résultats d'un essai consiste à affirmer un effet positif (efficacité ou sûreté) du traitement étudié plus grand que celui prouvé dans l'essai (cf. Table 1) (Boutron et al., 2010 ; Boutron et al., 2014 ; Haneef et

al., 2015 ; Yavchitz et al., 2016). Le phénomène des affirmations inappropriées est appelé « spin » dans la littérature anglaise sur le sujet, et nous allons utiliser ce nom dans ce qui suit[1].

| Résumé avec affirmation inappropriée | Résumé réécrit |
|---|---|
| *Treatment A **may be useful** in controlling cancer-related fatigue **in patients who present with severe fatigue** but is not useful in patients with mild or moderate fatigue.* | *Treatment A **was not more effective** than placebo in controlling cancer-related fatigue.* |
| *This study confirmed a previous trial demonstrating **improved PFS and response** for the treatment A compared with comparator B alone, although this did not result in improved survival.* | *The treatment A **was not more effective** than comparator B on overall survival in patients with metastatic breast cancer previously treated with anthracycline and taxanes.* |

**Table 1** : Exemples des résumés avec des affirmations inappropriées et des résumés réécrits par les experts afin que les conclusions correspondent aux résultats de la recherche. Les exemples ont été fournis par Isabelle Boutron et font partie de l'appendice de l'article (Boutron et al., 2014).

Plusieurs types d'essais médicaux se prêtent particulièrement bien à l'utilisation de spin, comme par exemple, les essais randomisés contrôlés[2] ou essais non-randomisés ainsi que les études d'exactitude de diagnostique (Boutron et al., 2010 ; Boutron et al., 2014 ; Lazarus et al., 2015 ; Yavchitz et al., 2016). En particulier, le spin est estimé être présent à différents degrés dans 60% des essais randomisés contrôlés qui sont la source la plus importante d'information de la médecine fondée sur les faits - *Evidence-Based Medicine* - (Boutron et al., 2014).

Le spin peut se trouver dans le résumé ainsi que dans le texte principal d'un article, mais il est plus fréquent dans les résumés (Boutron et al., 2010). Dans (Boutron et al., 2014), les auteurs ont étudié des articles du domaine de l'oncologie et ont prouvé que la présence de spin dans le résumé a de l'influence sur la manière dont les médecins interprètent l'effet positif du traitement discuté. Plus précisément, les médecins qui ont lu un résumé contenant du spin accordent plus d'efficacité au traitement que celle justifié par le corps de l'article. Le spin induit aussi une présentation inappropriée des résultats de recherche dans les communiqués de presse et les actualités de santé (Yavchitz et al., 2012 ; Haneef et al., 2015). Parfois le résumé est la seule partie disponible d'un article, augmentant ainsi fortement l'impact du spin sur la diffusion des résultats de recherche. La présence de spin influence les décisions prises par les médecins concernant l'utilisation des médicaments, le spin pose donc un problème grave, d'autant plus qu'il peut aussi affecter la perception que les patients ont d'un traitement. Il y a donc nécessité d'assister les auteurs d'articles scientifiques et leurs relecteurs à identifier les occurrences probables de spin. L'hypothèse explorée par notre projet est que les méthodes du Traitement Automatique des Langues (TAL) peuvent être utilisées pour détecter le spin en identifiant les affirmations inappropriées. Notre objectif est de développer des méthodes d'extraction des affirmations importantes et des éléments de preuve associés. Sur la base de la correspondance entre les affirmations et leurs informations justificatives nous allons calculer des traits caractéristiques qui seront utilisés par l'apprentissage automatique pour l'identification des textes contenant du spin. Dans ce but, nous allons collecter un corpus de textes biomédicaux et annoter les affirmations importantes, l'information justificative et l'absence/présence de spin.

[1] En français le terme « spin » est aussi utilisé, moins fréquemment on parle d'étude « embellie » ou encore de traitement « beautifié » cf. http://www.h2mw.eu/redactionmedicale/spin/

[2] Un essai randomisé contrôlé est un type d'étude scientifique utilisé pour tester l'efficacité d'un traitement. Les patients éligibles sont répartis au hasard en plusieurs groupes semblables : un (des) groupe(s) recevant le(s) traitement(s) étudié(s) et un groupe recevant un traitement de comparaison.

Dans cet article nous présentons la phase initiale du projet. Dans la section 2 nous présentons la notion de spin plus en détail, nous définissons les caractéristiques linguistiques des textes contenant du spin et examinons les fonctions de TAL que nous considérons comme nécessaires à son identification automatique. Dans la section 3 nous présentons les recherches en TAL qui ont abordé des tâches comparables à l'identification du spin et montrons comment nous comptons déployer les approches du TAL dans notre futur détecteur de spin. Dans la section 4 nous présentons nos recherches actuelles pour la tâche d'extraction des entités. Pour finir, dans la section 5 une conclusion et quelques perspectives seront présentées.

# 2 Notion de spin

Les études sur le spin ont distingué plusieurs types de spin, dont la fréquence varie en fonction du type d'essai médical abordé. Nous nous concentrons ici sur le spin dans les essais randomisés contrôlés (« *randomized controlled trials* », ou « *RCTs* »). Cette décision est basée sur l'importance de ce type d'essai par rapport aux autres types et est en accord avec les études existantes sur le spin. Dans ces dernières (Boutron et al., 2010 ; Lazarus et al., 2015 ; Yavchitz et al., 2016), les affirmations inappropriées sont classifiées en trois catégories et plusieurs sous-catégories :

1. présentation inappropriée des résultats de la recherche :
   - les effets négatifs ne sont pas présentés ;
   - on ne présente pas tous les résultats évalués (le résultat primaire n'est pas présenté ; on se concentre sur les résultats secondaires qui sont statistiquement significatifs)
   - la présentation du type et des caractéristiques de l'essai est imprécise ;
   - la présentation de la population étudiée est imprécise (on se concentre sur l'analyse de sous-groupes pour lesquels les résultats ont une significativité statistique ; on présente les résultats pour la population modifiée, par exemple après exclusion d'une partie des patients)
   - spin linguistique (utilisation des mots évaluatifs positifs de haut degré, comme *excellent*) ;
   - les limitations de l'essai ne sont pas présentées ;
   - les études précédentes sont partiellement citées (des articles importants ne sont pas cités) ;
2. interprétation inappropriée des résultats de la recherche :
   - on affirme que le médicament étudié a un effet positif ou l'effet équivalent au traitement de référence, bien que les résultats n'aient pas de significativité statistique ;
   - on présente le traitement comme sûr, bien que les résultats concernant la sûreté n'ont pas de significativité statistique ;
   - on affirme que le médicament étudié a un effet positif sans avoir fait de tests comparatifs ;
   - on considère la significativité statistique au lieu de la pertinence clinique ;
   - on affirme qu'il y a un effet causal entre le traitement étudié et le résultat évalué, bien que l'essai ne soit pas randomisé ;
3. extrapolation inappropriée :
   - au lieu de la population, l'intervention ou le résultat évalué, on présente une population plus large, une intervention ou un résultat différent ;
   - les implications tirées des résultats sont inappropriées pour la pratique clinique (conseils d'utiliser le traitement)

Le phénomène de spin est compliqué et hétérogène. Nous nous concentrons dans ce projet sur les types de spin identifiables uniquement à partir du texte de l'article, sans recourir à d'autres sources, par exemple les protocoles des recherches. Sur la base des types de spin, nous avons compilé une liste des tâches issues du domaine de TAL qui feront partie de l'identification automatique du spin :

1. extraction de l'évaluation du traitement (positif/neutre/négatif) ;
2. classification des articles biomédicaux selon le type d'essai (à ce moment nous nous sommes le plus intéressés à la distinction entre les essais randomisés contrôlés et les autres) ;
3. analyse de la structure du texte (division entre les parties titre/résumé/corps de l'article) ;
4. extraction des entités :
   - extraction des résultats étudiés, de préférence en distinguant le résultat primaire (*primary outcome*) des résultats secondaires (*secondary outcome*)[3] ;
   - extraction de la population et des groupes de patients ;
   - extraction de la significativité statistique des résultats ;
   - extraction des restrictions de l'essai ;
   - extraction des effets négatifs ;
   - extraction du traitement étudié et du traitement de référence.
5. extraction des relations entre les entités (par exemple relation entre résultats et significativité statistique)
6. analyse des paraphrases :
   - le résultat principal déclaré dans le corps de l'article est-il présenté dans le résumé ?
   - la population présentée dans le résumé est-elle la même que la population étudiée déclarée dans le corps de l'article, ou est-elle plus large (extrapolation), ou plus petite (sous-groupe) ?
7. analyse syntaxique : identification des constructions spécifiques au spin, par exemple proposition concessive associée souvent avec un changement de focus :

*« Treatment A may be useful in controlling cancer-related fatigue in patients with severe fatigue **but** is not useful in patients with mild or moderate fatigue »* (focus sur sous-groupe de patients).

*« This study demonstrates improved PFS and response for the treatment A compared with comparator B, **although** this did not result in improved survival »* (focus sur résultats secondaires).

# 3 Etat des recherches

Dans cette section nous allons considérer l'état des recherches en TAL pour deux tâches : 1) l'évaluation du biais, qui est lié à la détection de spin ; 2) l'extraction des entités et des relations entres elles, qui sont nécessaires à l'identification du spin.

## 3.1 Évaluation du biais (*bias assessment*)

Dans la littérature biomédicale, la tâche la plus proche de l'identification du spin est l'évaluation du biais dans les études cliniques. Le biais est défini comme une erreur systématique ou une déviation par rapport à la réalité dans les résultats ou les conclusions, qui peut induire une sous-estimation ou une surestimation de l'effet du traitement examiné (Higgins and Green, eds, 2008). Les erreurs peuvent concerner la conception de l'étude (*study design*), le déroulement de la recherche, l'analyse et la présentation des résultats. Les types de biais adressés incluent :

- le biais de sélection (*selection bias*) : génération de la séquence aléatoire ; masquage de l'allocation du traitement;
- le biais de performance (*performance bias*) : anonymisation (ou insu[4)] des participants et du personnel ;

[3] Les résultats (*« outcomes »)* sont les variables mesurées dans un essai. Elles doivent être déterminées à l'avance et divisées en primaires (les plus importantes pour prouver l'effet favorable ou défavorable du traitement) et secondaires (utilisées pour estimer les effets supplémentaires).

– le biais de détection (*detection bias*) : anonymisation de l'évaluation des résultats ;
– le biais d'attrition (*attrition bias*) : données incomplètes sur les résultats ;
– le biais de présentation (*reporting bias*) : présentation sélective des résultats.

Le biais de présentation nous intéresse le plus parce qu'il tombe aussi sous la notion du spin.

Plusieurs outils existent pour l'évaluation du risque de biais. Selon (Higgins et al., 2011), l'évaluation est le plus souvent faite par des experts en utilisant les méthodes autre que le TAL : des échelles (*scales*) et des listes de contrôle (*checklists*). Il y a aussi des approches qui explorent les méthodes de TAL pour automatiser l'évaluation de biais (Marshall et al., 2015-1, Marshall et al., 2015-2). Dans (Marshall et al., 2015-1), les auteurs décrivent la création d'un corpus sur la base des revues systématiques archivées et mises à disposition par le réseau Cochrane (*Cochrane Database of Systematic Reviews*, *CDSR*[5]). Cochrane[6] est un réseau international indépendant de chercheurs, professionnels de santé et patients dont le but d'améliorer la prise de décision dans le domaine de la santé. Les revues systématiques du CDSR contiennent l'information structurée (y compris l'évaluation de biais) sur les études, et c'est cette information qui est utilisée comme annotation du corpus. Les auteurs remarquent que l'évaluation de biais est subjective ; l'accord entre les experts étant le plus bas pour le biais de présentation. Il faut noter aussi que l'évaluation du biais de présentation est à la base des protocoles d'études. Les auteurs utilisent le modèle de machine à vecteurs support (« *Support Vector Machine* ») pour la classification des articles selon le niveau de biais, en prenant les mots du texte comme caractéristiques (approche sac de mots). Le système est présenté plus en détails dans (Marshall et al., 2015-2) où les auteurs comparent différents modèles pour l'évaluation de biais. Les résultats montrent que les données pour les biais d'attrition et de présentation sont très bruitées. Le modèle proposé par les auteurs a de meilleures performances pour les autres types de biais lorsque les données des biais d'attrition et de présentation sont exclues.

En conclusion, la tâche d'évaluation de biais est liée à la détection de spin car les deux regardent le problème de la présentation sélective. Les méthodes de TAL ont été utilisées pour l'évaluation de biais. Les recherches ont montré que l'accord entre les experts dans ce domaine est bas est que les données associées à certaines catégories de biais sont très bruitées, ce qui est problématique pour l'apprentissage automatique. La différence entre les deux tâches est que l'évaluation de biais considère la correspondance entre le protocole de la recherche et la présentation des résultats, tandis que nous n'analysons pas les protocoles pour le moment. Nous considérons la correspondance entre les caractéristiques d'une étude décrite dans le corps d'un article et celles présentées dans le résumé.

## 3.2 Extraction des entités et des relations

Le volume de recherche en extraction des entités de textes biomédicaux est assez grand, mais la plupart de la recherche est concentrée sur l'extraction des noms des gènes, protéines et médicaments. L'extraction des entités importantes pour nous a moins attiré l'attention (Summerscales et al., 2011).

---

[4] L'anonymisation (ou insu, par exemple on parle de « levée d'insu ») est utilisé dans les essais médicaux comparant plusieurs médicaments (des traitements expérimentaux et des traitements de référence). On s'assure que les patients ou les examinateurs, ou les deux, ne savent pas qui reçoit quel traitement. Comme pour l'évaluation d'articles scientifique ou parle alors de tests en « simple aveugle », « double aveugle » ou moins fréquemment de tests en « simple insu » ou « double insu » (http://www.spc.univ-lyon1.fr/polycop/double%20aveugle.htm).

[5] Cochrane Database of Systematic Reviews (CDSR) est une ressource primaire pour les revues systématiques dans le domaine médical. CDSR inclut des revues systématiques (les travaux dont le but est la synthèse des connaissances sur un sujet donné), les protocoles des revues et les éditoriaux.

[6] http://www.cochrane.org

Quand on parle des entités ici, on sous-entend les entités nommées (les noms propres comme les noms des médicaments ou des maladies), mais aussi les phrases (nominales, verbales ou autres) qui remplissent des « slots » (rôles sémantiques) dans le cadre sémantique décrivant les essais randomisés contrôlés : les résultats examinés, la population de patients, les effets négatifs secondaires, etc. (en s'inspirant de travaux comme Kiritchenko et al.,2010, Nguyen et al., 2013).

Cependant, la tâche d'extraction des entités concernant les caractéristiques des essais ou les descriptions des situations cliniques (y compris les informations concernant les patients, les traitements, les maladies ; les paramètres de l'essai, comme les résultats en question, les branches de l'intervention[7], les effets négatifs de l'intervention, etc.) a attiré de l'attention car elle est requise pour de nombreuses applications : résumé automatisé des textes, construction des revues systématiques, systèmes d'aide à la décision clinique, systèmes de réponse à des questions, création des bases de données structurées, recherche dans les bases de données, etc. Les approches utilisées pour l'extraction des entités varient en fonction du domaine et de la tâche spécifique abordée.

Pour la construction des revues systématiques, il faut trouver dans les articles sur les essais randomisés contrôlés au moins quatre éléments de base pour les études cliniques : population/problème, traitement, traitement de comparaison, et résultat (ces 4 éléments sont connus sous l'acronyme «*PICO* framework»[8] pour « *Population/Problem, Intervention, Comparator, Outcome »*). Mais il ne s'agit pas toujours de l'extraction des entités, car il suffit de trouver des phrases qui contiennent l'information recherchée sous une autre forme (Wallace et al., 2016).

(Dawes et al., 2007) regardent un ensemble plus large d'éléments : « *Patient–Population–Problem, Exposure–Intervention, Comparison, Outcome, Duration and Results* (*PECODR*) ». Les auteurs se concentrent sur les résumés. La tâche de cette recherche n'est pas encore de créer un système d'extraction des entités, mais d'explorer la possibilité d'extraire automatiquement ces types d'entités. Dans ce but, les auteurs retrouvent les termes qui sont associés à chaque élément (les contextes possibles comme « *outcome »* ou « *end point »* pour le résultat (*Outcome*), et aussi les mots qui référent directement à l'entité, comme « *mortality »* pour l'entité *Outcome*). Le nombre de résumés étudiés est très petit (20 résumés). Sur la base de ce corpus, ils font la conclusion qu'il y a des mots spécifiques associés avec certains éléments (« *Comparison »*, « *Outcome »* et « *Results »*) qui sont utilisés dans plusieurs articles, alors que les autres éléments (« *Patient–Problem »* et « *Exposure »*) ne sont pas définis par un ensemble de mots communs pour un grand nombre d'articles.

(De Bruijn et al., 2008) élargissent encore l'ensemble d'éléments, leur objectif est d'extraire des textes des essais randomisés contrôlés tous les éléments du *«Standards fusionnés dans la rédaction d'essais thérapeutiques» («CONSORT statement»*[9]) comme : les critères d'éligibilité, le nom du traitement étudié et le traitement de comparaison, les paramètres de l'intervention, y compris le dosage, la fréquence, la durée, etc., la taille de l'échantillon, la date de début et de fin d'inscription, les résultats primaires et secondaires avec les indicateurs temporels associés; l'information sur le financement, les métadonnées de la publication, y compris la date, les auteurs. L'objectif général est

[7] Les descriptions des branches de l'intervention comprennent normalement le nom du traitement expérimental et du traitement de référence, elles peuvent aussi inclure le dosage, les horaires et autres détails concernant l'administration.

[8] https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0029906/

[9] Le « Consolidated Standards of Reporting Trials » ou « CONSORT group » est un groupe regroupant des experts en méthodologie d'essais cliniques, experts en bonnes pratiques cliniques, des éditeurs de journaux du domaine biomédical et de sponsors de la recherche biomédicale dont le but est d'améliorer la qualité de la rédaction des essais randomisés contrôlés. Le «CONSORT Statement» est une liste d'items à vérifier pour évaluer la rédaction d'un essai. Site officiel: http://www.consort-statement.org/

d'extraire l'information essentielle sur les études qui puisse être utilisée dans des bases de données structurées. L'algorithme procède en deux étapes. D'abord, on utilise un classificateur pour choisir 5 phrases les plus pertinentes pour chaque élément. Ensuite, on applique des règles pour l'extraction des entités des phrases choisies. Les auteurs ont choisi comme mesure d'évaluation la correspondance partielle entre les séquences des mots obtenues et l'entité ciblée. La coïncidence totale entre les séquences trouvées et les entités n'est pas évaluée car déterminer les limites précises des entités est une tâche difficile. Pour certains types d'information (auteurs, DOI, date de publication), ce sont les métadonnées de la base de citation et de résumés d'articles du domaine biomédical Medline qui sont utilisées.

Ce système se distingue de la majorité des autres approches par le fait qu'il travaille avec des articles entiers au lieu des seuls résumés. L'approche a été développée dans les recherches suivantes des mêmes auteurs (Kiritchenko et al., 2010) qui ont créé le système appelé ExaCT pour l'extraction des entités des textes des essais randomisés contrôlés. Le système a pour objectif d'aider les examinateurs à identifier l'information nécessaire mais n'a pas pour objectif l'automatisation complète de l'extraction exhaustive de l'information.

Plusieurs études concernent l'extraction des entités pour le résumé automatisé de textes (Summerscales et al., 2009 ; Summerscales et al., 2011). Dans le premier les auteurs présentent les résultats préliminaires d'application des méthodes d'extraction des entités aux résumés des articles sur les études contrôlées du BMJ[10]. L'objectif est de trouver les noms des traitements, les groupes expérimentaux et les résultats examinés. Les auteurs décrivent quelques difficultés liées à ces types d'entités, telles que l'hétérogénéité des entités ciblées (une entité peut être représentée par un nom simple, par une phrase complexe, par une ellipse ; les traitements et résultats manquent de caractéristiques orthographiques spécifiques comme des chiffres, des lettres majuscules, ou des caractères spéciaux). Les auteurs utilisent le modèle de champs aléatoires conditionnels (« C*onditional Random Fields » ou «CRF»*) pour la classification des entités. Ils font des comparaisons entre différents ensembles de caractéristiques utilisées pour la classification: le mot ; sa catégorie lexicale ; l'identifiant de thématique médicale (« *Medical Subject Heading Id »*) associé au mot ; des étiquettes sémantiques (ici les types sémantiques des mots définis dans le métathesaurus UMLS[11], qui sont trouvés avec l'aide du programme MetaMap (Aronson, 2001) ; le titre de la section du résumé où le mot apparaît ; les quadrigrammes gauche et droite avec leurs catégories lexicales et leurs étiquettes sémantiques. Les auteurs font la conclusion que les caractéristiques qui comprennent les catégories thématiques et sémantiques du mot sont plus utiles que celles basées sur sa forme, notamment pour l'extraction des résultats examinés. Les auteurs supposent que l'utilisation des caractéristiques syntaxiques pourrait améliorer la performance de l'algorithme.

Dans (Summerscales et al., 2011) les auteurs présentent un système pour le calcul automatique de statistiques sommaires pour les essais randomisés contrôlés, y compris des valeurs de réduction de risque absolu (*Absolute Risk Reduction*, *ARR*)[12] et du nombre de sujets à traiter[13]. A cette fin, il est

[10] Le « British Medical Journal » (appelé officiellement « BMJ » depuis 1988 et « The BMJ » depuis 2014) est une revue britannique de médecine générale, l'une des revues médicales les plus anciennes et les plus lues dans le monde. Site officiel: http://www.bmj.com/

[11] L'UMLS (Unified Medical Language System) est un compendium des plusieurs vocabulaires contrôlés médicaux. Site officiel: https://www.nlm.nih.gov/research/umls/

[12] La réduction du risque absolu est la différence entre la probabilité de survenue d'un événement dans le groupe expérimental et la probabilité de survenue de cet événement dans le groupe témoin (groupe qui reçoit le traitement de référence).

[13] Le nombre de sujets à traiter est le nombre de patients qu'il est nécessaire de traiter pendant une période donnée pour éviter 1 événement défavorable.

nécessaire d'extraire les descriptions et les tailles des groupes de patients, les résultats étudiés et le nombre de résultats positifs/négatifs. D'abord l'algorithme cherche à identifier les phrases contenant l'information pertinente. Ensuite, un classificateur à base de CRF est appliqué pour trouver les mentions des groupes expérimentaux, des résultats examinés, la taille des groupes expérimentaux et le nombre de résultats positifs/négatifs. L'algorithme cherche ensuite à établir des relations entre les entités (taille du groupe + définition du groupe ; résultat étudié + nombre de résultats positifs/négatifs) et à calculer des statistiques sommaires sur la base des informations extraites. Cette recherche montre que les résultats examinés (« outcomes ») sont plus difficiles à extraire que les autres types d'informations.

(Chung, 2009) se concentre elle, sur l'extraction des « branches d'intervention », c-a-d les différents traitements ou placebos mis en jeux lors de l'essai, souvent associés chacun à un groupe de patients, des essais randomisés contrôlés. L'auteur explore la possibilité d'extraire les branches de l'intervention sur la base de l'analyse des constructions coordonnées. L'approche est basée sur l'analyse des phrases de la section « Méthodes » des résumés où l'information concernant les branches est normalement présentées explicitement ; les sections « Objectifs », « Résultats », et « Conclusions », où l'information pertinente peut apparaître implicitement, ne sont pas concernées. L'auteur a collecté et analysé un corpus d'essais randomisés contrôlés et a abouti à la conclusion que les constructions coordonnées sont fréquentes dans les descriptions de branches. L'auteur explore l'hypothèse que l'analyse syntaxique automatique des phrases complètes puisse être utilisée pour trouver les constructions coordonnées qui décrivent les branches de l'intervention. L'algorithme comprend deux étapes, d''abord les phrases du résumé sont classées pour identifier celles qui appartiennent à la section méthodes. Ensuite, les phrases sont normalisées, et un classifieur basé sur l'entropie maximale est utilisé avec des ensembles de caractéristiques différents (conjonction de coordination ; type de coordination ; arguments syntaxiques, etc.). Les classes sémantiques des mots sont affectées sur la base de l'UMLS avec l'aide de MetaMap.

D'autres recherches se sont intéressées à l'extraction des informations concernant la population des patients. Par exemple, (Xu et al., 2007) proposent une approche pour extraire les descriptions de population, le nombre de patients examinés et les descriptions des maladies/symptômes des essais randomisés contrôlés. L'algorithme procède en deux étapes. La première consiste à rechercher des phrases qui puissent contenir l'information sur la population ; un modèle de Markov caché est utilisé. La deuxième étape concerne l'extraction des entités avec des règles syntaxiques. Une approche similaire est explorée dans (Raja et al., 2016), les auteurs utilisent un classifieur binaire pour trouver les phrases contenant l'information sur la population ; ensuite ils appliquent des règles utilisant des expressions régulières et une analyse syntaxique pour extraire les entités.

De ces travaux antérieurs, nous pouvons tirer quelques conclusions :

- la plupart des recherches sont concentrées sur les essais randomisés contrôlés ;
- la plupart des recherches considèrent l'extraction des entités seulement à partir des résumés ;
- l'approche commune comporte deux étapes : filtrer les phrases qui puissent contenir les entités ; trouver les entités dans les phrases filtrées.
- la plupart des approches combinent l'apprentissage automatique et les règles ;
- Des systèmes comme MetaMap pour associer les termes utilisés dans les articles aux termes de l'UMLS sont utilisés souvent ;
- définir les limites précises d'une entité est une tâche plus difficile qu'obtenir une correspondance partielle ;
- le résultat examiné (*outcome*) est une des entités les plus difficiles à identifier.

# 4 Expériences

## 4.1 Le corpus

Nous allons collecter un corpus pour la détection du spin et l'annoter avec l'information pertinente plus tard au cours de notre projet. En ce moment, nous utilisons pour l'analyse préliminaire un corpus d'articles de PMC[14] collecté lors d'une expérience précédente effectuée au laboratoire. Nous avons identifié 3938 articles concernant des essais randomisés contrôlés dans notre corpus sur un total de 65396 articles.

## 4.2 Extraction des entités nommées : résultat examiné (*outcome*)

L'entité la plus importante pour l'identification de spin est le résultat examiné (*outcome*) parce que plusieurs types de spin concernent ce type d'entité. Nous allons utiliser le mot anglais *outcome* dans cette section pour ne pas confondre le nom de l'entité et le mot *résultat* avec son sens usuel.

Nous avons mentionné plusieurs études qui abordent la tâche d'extraction de l' « outcome ». Ces études considèrent les phrases comme l'exemple suivant qui provient de l'article (Summerscales et al., 2009) : «*Mortality was higher in the quinine than in the artemether»*. Les auteurs remarquent que sauf pendant la présentation des résultats, l' « outcome » est normalement mentionné explicitement dans le résumé ou dans le corps de l'article ou bien dans les deux, ceci avec des expressions du type :

- *Expected outcome is… (« le résultat attendu est ... »)*
- *... was our primary outcome. (« ... était notre résultat principal »)*
- *Outcomes include...(« le résultat inclut ... »)*

Les études citées ci-dessus ne cherchent pas à identifier ce genre d'expressions car leur objectif est d'extraire les résultats obtenus et présentés par les recherches. Notre but est différent : retrouver les descriptions des « outcomes » initiaux d'un essai ainsi que les « outcomes » présentés dans l'article étudié, pour identifier les cas où ils ne correspondraient pas. Nous pouvons donc nous concentrer en première approche sur les constructions données en exemple précédemment (« *outcome is* ... »), ce qui est une tâche plus simple que celles présentées dans l'état de l'art de la section 3.2.

Parmi les types de spin les plus fréquents, nous trouvons l'absence de l' « outcome » primaire dans la présentation des résultats et l'accent mis sur les « outcomes » secondaires. Nous nous sommes donc intéressés à la distinction entre les deux types d' « outcome » : primaire et secondaire. Actuellement notre but est d'évaluer la possibilité d'extraire les « outcomes » des descriptions explicites et d'identifier l' « outcome » primaire. Notre recherche est basée sur 3938 textes de notre corpus qui ont le type «*Randomized controlled trial »*. Nous procédons selon les étapes suivantes :

1. Nous avons identifié manuellement toutes les phrases du corpus qui contiennent des mots qui puissent être utilisés pour décrire des outcomes :
   - les mots « *outcome* », « *endpoint* » (avec ses variantes orthographiques « *end point* », « *end-point* ») qui décrivent directement l'entité « outcome » ;
   - les noms « *aim* », « *purpose* », « *objective* », etc., les verbes « *explore* », « *investigate* », « *examine* », etc., qui qui peuvent servir comme marqueurs de l'entité « *outcome* ».

[14] PMC (PubMed Central) est une base de données d'articles en génie biomédical et dans les sciences de la vie. Site officiel : https://www.ncbi.nlm.nih.gov/pmc/

2. Nous avons analysé les phrases ainsi obtenues pour trouver les constructions fréquentes utilisées pour décrire l' « outcome ». Nous avons porté une attention spécifique à deux points importants : a) la distinction entre la description la plus explicite des « outcomes » (avec les mots « *outcome* » et « *endpoint* ») et les autres variantes des descriptions de l'objectif ; b) la distinction entre l' « outcome » primaire et les « outcome » secondaires.
3. Nous avons créé des règles à base de grammaires locales avec la boîte à outils Unitex (Paumier, 2016) pour trouver ces constructions. La figure 1 donne un exemple de graphe pour identifier les « outcomes ». Actuellement nous utilisons 9 graphes pour détecter tous les « outcomes » possibles.

Nous cherchons aussi à séparer le résumé du corps de texte. Dans ce but, nous avons créé des règles qui réussissent à trouver les résumés pour 3935 articles de notre corpus.

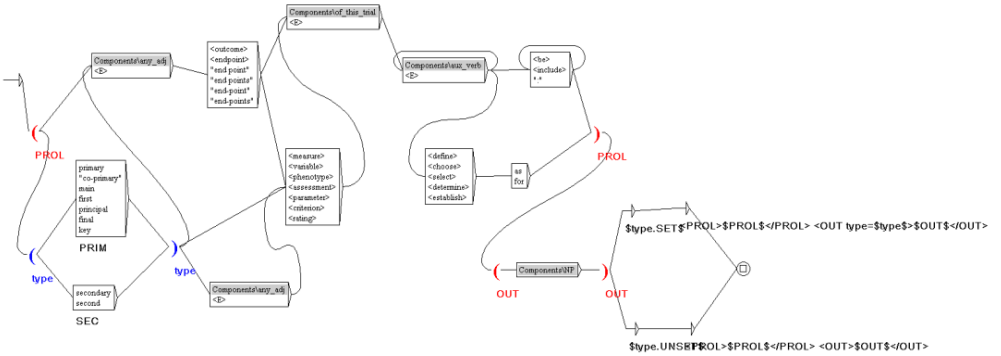



Figure 1 : Un exemple du graphe pour identifier les « outcomes »

Il y a plusieurs raisons pour lesquelles nous utilisons des règles pour la détection des parties des articles et pour l'extraction des entités. L'approche basée sur des règles est une alternative à l'apprentissage automatique pour ces tâches. Par exemple, CasSys et CasEn sont deux systèmes pour l'extraction des entités employant des règles à base de grammaires locales (Friburger et al., 2004 ; Maurel et al., 2011). L'autre raison pour employer les règles au stade actuel de notre projet est l'absence de corpus annoté pour notre recherche. La collecte d'un corpus des textes contenant du spin fait partie des étapes suivantes de notre travail. Dès que le corpus sera annoté avec l'information nécessaire pour identifier le spin, y compris les entités, nous pourrons employer des méthodes d'apprentissage automatique et comparer leur performance avec celle du système basé sur des règles. Le système actuel servira donc comme une approche de référence. En plus, nous comptons utiliser les grammaires locales pour une pré-annotation du corpus, avant annotation par les annotateurs humains. En absence de corpus annoté nous n'avons pas d'évaluations précises de la performance de l'algorithme actuel, les estimations préliminaires ont été effectuées sur un petit échantillon d'exemples compilés manuellement.

Un exemple des résultats que nous obtenons est montré dans la table 2, avec les notations suivantes :

– *Location* correspond à la partie du texte (*Abstr* - le résumé, *Body_text* - le texte principal).
– *Extracted words* sont les séquences des mots que nous avons identifiées comme candidates pour être un « outcome ».
– *Sentence* est une phrase où l'on a trouvé une occurrence d'une construction. <PROL> marque le contexte de l'entité candidate ; <OUT> marque l'entité.

| *Location* | *Extracted words* | *Sentence* |
|---|---|---|
| Abstr | the remission of depressive symptoms at the 2-month follow up visit | The <PROL>primary outcome was</PROL> <OUT type=PRIM>the remission of depressive symptoms at the 2-month follow up visit</OUT>, defined as a HDRS score of 7 or less. |
| Body_text | overall mortality, severity of BPD, number of days on the ventilator, number of treatment failures, ventilation-induced lung injury and pulmonary hypertension | <PROL>Secondary outcome parameters are</PROL> <OUT type=SEC>overall mortality, severity of BPD, number of days on the ventilator, number of treatment failures, ventilation-induced lung injury and pulmonary hypertension</OUT> according to clinical and laboratory parameters and need for ECMO (only for ECMO centres). |

**Table 2** : Résultat d'application des règles

## 4.3 Extraction des entités nommées: population/groupes des patients

Pour la tâche d'extraction de population/groupes des patients notre objectif est identique aux recherches précédentes. Pour le moment nous ne regardons pas spécialement l'extraction du nombre de participants, mais seulement les descriptions de patients (âge, sexe, nationalité). Nous avons utilisé la même approche que pour l' « outcome »: nous avons collecté toutes les phrases qui contiennent une liste des mots pertinents (« *adults* », « *children* », « *adolescents* », « *population* », « *cohort* », etc.). Ensuite nous avons analysé les constructions qui décrivent la population étudiée. La figure 2 donne un exemple du graphe utilisé pour identifier la population étudiée. Actuellement nous utilisons 5 graphes pour détecter les descriptions de la population.

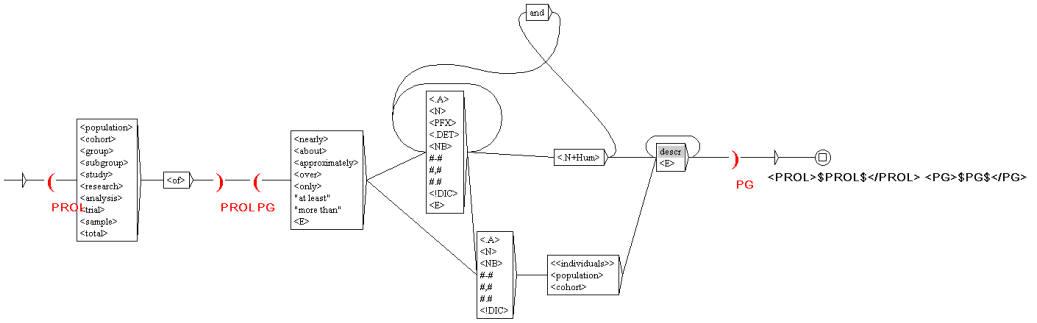



Figure 2 : Un exemple du graphe pour identifier la population étudiée

Pour faire la recherche la plus stricte possible, nous limitons les séquences des mots trouvées (notées par <...>) en exigeant qu'elles contiennent des mots du groupe sémantique « humains », que nous déterminons par une liste des mots compilée manuellement et par le trait <.N+HUM> des dictionnaires Unitex. Nous avons inclut aussi la construction plus générale qui contient des mots comme « *adults* », « *children* », « *adolescents* », etc. précédé ou suivis par une description (adjectif, nombre, groupe prépositionnel). Pour cette construction nous n'utilisons pas le trait <.N+HUM>.

Les constructions ci-dessus ont été trouvées dans 3935 textes (99,9%) de notre corpus. Mais la précision de la recherche est assez basse: notre analyse montre que les constructions étudiées sont souvent utilisées dans les rubriques des articles dédiées à la revue de la littérature, pour décrire la population examinée dans les études antérieures, ce que nous ne voulons pas extraire. Notre tâche future sera de créer des algorithmes pour filtrer les constructions trouvées.

# 5 Conclusion et perspectives

Dans ces premières expériences sur la détection du spin dans les articles scientifiques, nous avons examiné la possibilité d'extraire les résultats (« outcomes ») d'un essai randomisé contrôlé et la population ou les groupes des patients, avec l'aide de grammaires locales. Nous avons mis l'attention particulièrement sur le « primary outcome » (résultat primaire), qui est crucial pour la détection de spin, sur la base des constructions qui décrivent les objectifs et les résultats étudiés.

Nos résultats montrent que les constructions simples qui décrivent les « outcomes » et la population couvrent une grande partie de textes (99,9% pour la population, 91,5% et 94% pour les constructions les plus générales concernant l' « outcome » examiné). Un algorithme du filtrage automatique des résultats trouvés est cependant encore nécessaire pour identifier les entités ciblées.

Concernant les constructions plus spécifiques, une mention explicite du « primary outcome » est trouvée dans 51,8% d'articles, une mention de l'objectif principal (du type « *We aim at <...>* » / « *Our main goal is <...>* », etc.) est présente dans 21,1% des articles. La construction « *<...> is assessed/investigated/studied/analyzed/explored* » est utilisée pour décrire des outcomes, mais aussi les groupes de patients et toutes les mesures utilisées dans une étude. Elle est la plus fréquente (apparaît dans 94% des textes). Nous supposons qu'en l'absence d'une mention explicite d' « outcome », on peut identifier le « primary outcome » sur la base des mentions de l'objectif principal et des mesures utilisées. Cependant, cette hypothèse doit être discutée avec les experts et vérifiée sur un corpus annoté. Une analyse préliminaire montre qu'afin d'extraire l' « outcome » des constructions générales, il faudrait développer des algorithmes supplémentaires, tel que:

- des règles de filtrage pour éliminer les phrases des types « *Outcomes were measured* » et « *Study subjects were evaluated every two weeks* ». Dans la recherche future, nous allons tester des règles simples basées sur les mots indésirables (*outcome*, *variable*, etc.) et l'intersection avec l'entité *Patient group/Population* ;
- des règles d'extraction des mentions de l'objectif lorsque les mentions sont très générales (e.g. « *Our aim is* ***to evaluate efficacy*** » dont on doit extraire « efficacy »).

Nos prochains travaux concerneront l'extraction de l' « outcome » à partir de références implicites, l'association des « outcomes » avec leur significativité statistique, et la vérification de la correspondance entre les « outcomes » décrits dans le corps d'un article et dans les sections « Résultats » et « Conclusions ». Nous sommes intéressés par les phrases du type suivant : « ***Survival rate*** *was higher in group A* ***(p=0.0003)*** *».* Etant donné la difficulté d'extraire l' « outcome » dans de telles phrases, nous allons explorer la possibilité d'extraire plusieurs phrases candidates. Nous comptons aussi investiguer l'emploi de traitement spécifique lorsque les rôles sémantiques sont remplis par des entités nommées et bien sûr approfondir notre modèle de rôles sémantiques par rapport aux modèles utilisés pour l'extraction d'évènements, les « frames », etc.

# Remerciements

Ces travaux ont été effectués dans le cadre du projet MIROR financé par le programme de recherche et d'innovation de l'Union Européenne Horizon 2020 sous la référence « Marie Sklodowska-Curie grant agreement No 676207 ».

# Références

ARONSON A. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. Proc. *AMIA Symposium*.

BOUTRON I., ALTMAN D.G., HOPEWELL S., VERA-BADILLO F., TANNOCK I., RAVAUD P. (2014). Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of Cancer: the SPIIN randomized controlled trial. *J Clin Oncol*, 32, 4120–4126.

BOUTRON I., DUTTON S., RAVAUD P., ALTMAN D.G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 303, 2058–2064.

CHUNG G.Y. (2009). Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *J Biomed Inform*, 42(5), 790-800.

DAWES M., PLUYE P., SHEA L., GRAD R., GREENBERG A., NIE J.-N. (2007). The identification of clinically important elements within medical journal abstracts: Patient-population-problem, exposure-intervention, comparison, outcome, duration and results (PECODR). *Informatics in Primary Care*, 15(1), 9–16.

DE BRUIJN B., CARINI S., KIRITCHENKO S., MARTIN J., SIM I. (2008). Automated information extraction of key trial design elements from clinical trial publications. Proceedings of *the AMIA Annual Symposium*, 141-145.HANEEF R., LAZARUS C., RAVAUD P., YAVCHITZ A., BOUTRON I. (2015). Interpretation of results of studies evaluating an intervention highlighted in Google Health News: a cross-sectional study of news. *PLoS ONE*, 10(10), doi:10.1371/journal.pone.0140889.

FRIBURGER N., MAUREL D. (2004). Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, 313, 94-104.

HIGGINS J.P., ALTMAN D.G., GOTZSCHE P.C., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343:d5928

HIGGINS J.P., GREEN S., eds. (2008). Cochrane handbook for systematic reviews of interventions. West Sussex : Wiley & Sons Ltd.

KIRITCHENKO S., DE BRUIJN B., CARINI S., MARTIN J., SIM I. (2010). ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 10: 56-10.1186/1472-6947-10-56.

LAZARUS C., HANEEF R., RAVAUD P., BOUTRON I. (2015). Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 15:85.

MARSHALL I.J., KUIPER J., WALLACE B.C. (2015-1). Automating risk of bias assessment for clinical trials. *IEEE journal of biomedical and health informatics*, 19(4), 1406-1412.

MARSHALL I.J., KUIPER J., WALLACE B.C. (2015-2). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, ocv044.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D. (2011). Cascades autour de la reconnaissance des entités nommées. *TAL* 52-1.

NGUYEN N., MIWA M., TSURUOKA Y., TOJO S. (2013). Open information extraction from biomedical literature using predicate-argument structure patterns. *Proceedings of the 5th International Symposium on Languages in Biology and Medicine,* 51–55.

PAUMIER S. (2016). Unitex 3.1 User Manual. http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf

RAJA K., DASOT N., TECH B., GOYAL P., JONNALAGADDA S.R. (2016). Towards evidence-based precision medicine: extracting population information from biomedical text using binary classifiers and syntactic patterns. *AMIA Jt Summits Transl Sci Proc*, 203-212.

SUMMERSCALES R.L., ARGAMON S., BAI S., HUPERFF J., SCHWARTZFF A. (2011). Automatic summarization of results from clinical trials. *The 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 372–377.

SUMMERSCALES R.L., ARGAMON S., HUPERT J., SCHWARTZ A. (2009). Identifying treatments, groups, and outcomes in medical abstracts. *The Sixth Midwest Computational Linguistics Colloquium* (MCLC 2009). 2009.

WALLACE B.C., KUIPER J., SHARMA A., ZHU M., MARSHALL I.J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132), 1–25.

XU R., GARTEN Y., SUPEKAR K.S., DAS A.K., ALTMAN R.B., GARBER A.M. (2007). Extracting subject demographic information from abstracts of randomized clinical trial reports. *Proceedings of the 12th World Congress on Health (Medical) Informatics*. Brisbane, Australia. Edited by: Kuhn K.A., Warren J.R., Leong T.Y. Amsterdam: IOS Press ; 2007:550-554.

YAVCHITZ A., BOUTRON I., BAFETA A., MARROUN I., CHARLES P., MANTZ J., et al. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 9:e1001308.

YAVCHITZ A., RAVAUD P., ALTMAN D.G., MOHER D., HROBJARTSSON A., LASSERSON T., BOUTRON I. (2016). A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of Clinical Epidemiology*, 75, 56-65.

# Finding Missing Categories in Incomplete Utterances *

Mehdi Mirzapour
LIRMM, Université de Montpellier, Montpellier, France
mehdi.mirzapour@lirmm.fr

RÉSUMÉ

Cet article propose un algorithme efficace (en $O(n^4)$) pour trouver la catégorie d'un mot manquant dans un énoncé incomplet. Notre travail fait appel à l'algorithme d'unification comme lors de l'apprentissage des grammaires catégorielles et à la programmation dynamique comme dans l'algorithme Cocke-Younger-Kasami. En utilisant l'interface syntaxique / sémantique des grammaires catégorielles, ce travail peut être utilisé pour dériver les lectures sémantiques possibles d'un énoncé incomplet. Des exemples suivis illustrent notre propos.

ABSTRACT

**Finding Missing Categories in Incomplete Utterances**

This paper introduces an efficient algorithm ($O(n^4)$) for finding a missing category in an incomplete utterance by using unification technique as when learning categorial grammars, and dynamic programming as in Cocke–Younger–Kasami algorithm. Using syntax/semantic interface of categorial grammar, this work can be used for deriving possible semantic readings of an incomplete utterance. The paper illustrates the problem with running examples.

MOTS-CLÉS : Syntaxe, Grammaires Catégorielles, Inférence grammaticale, Apprentissage, Unification, Programmation Dynamique..

KEYWORDS: Syntax, Categorial Grammars, Grammar Inference, Learning, Unification, Dynamic Programming..

## 1 Introduction

This paper is a part of the author's Ph.D. proposal regarding possible semantic readings of a given utterance with ranking. An utterance may yield several readings either from ambiguities (lexical, syntactical and semantical) or from syntactic incompleteness. The incomplete utterances can be linguistically considered as a type of non-canonical utterances (Pullum & Scholz, 2001) and they can be caused by different reasons such as missing categories, extra categories, swap categories, misused categories, or any possible combination of mentioned reasons. The focus in this paper is on incomplete utterances with one missing category. To make our research problem clear we should emphasize that by the missing category one can generally suppose two things : a missed word in an incomplete sentence or an unknown word in a sentence with no proper category assignment. In this paper we especially deal with the problem of a missing category for a missing word in an incomplete sentence. Moreover, we assume that the position of the word in the sentence is unknown. So obviously,

∗. This work is thought as a possible complement to the e-fran project AREN (ARgumentation Et Numérique) as a preprocessing technique for the automated analysis of online debates by high-school pupils.

our task is to find the possible position(s) of a missing word and its proper category assignment.

Syntactic analysis of incomplete utterances has been the subject of study in rule-based, data-driven and statistical approaches in computational linguistics (Leacock *et al.*, 2010). The scope of automatic error detection/fixation and robust parsing is so wide and different mechanisms are introduced by researchers such as [1] : over-generating and ranking parse trees (Dini & Malnati, 1993), imposing ranking constraints on grammatical rule Mellish (1989), introducing "mal-rules" (Schneider & McCoy, 1998), relaxing constraints in feature unification (Vogel & Cooper, 1995), modeling grammar as model-theoretic constraints (Blache, 2000; Dahl & Blache, 2004; Prost, 2009), and finally, sub-tree parse fitting for non-parsable utterances. (Jensen *et al.*, 1983). All the mentioned approaches have their own advantages, and address the same issue that we have defined. However, the formalisms that are introduced in them is not the same as what is used in this paper.

Basic Categorial Grammars or AB grammars, is a formalism which is used in this research. It is a family of the categorial grammars. This choice is based on the requirement to bridge the incomplete utterances with semantical readings (author's Ph.D. proposal). There are a number of reasons to explain why categorial grammars are suitable for our specific research topic : Firstly, categorial grammars have a very nice correspondence with semantics and they are very suitable for syntactic-semantic interface modeling ; they can represent semantic ambiguities such as quantifier scope and negation (Moot & Retoré, 2012, Chapter 3) which does not exist in shallow approaches ; secondly, they enjoy learn-ability property from positive sentences which converge after learning a fixed number of sentences by using unification techniques (Buszkowski & Penn, 1990; Kanazawa, 1998) ; thirdly, by applying a left-to-right incremental procedure and proof net construction for categorial grammar we can correctly predict a wide variety of human performance phenomena such as garden-pathing, preference for lower attachment, left-to-right quantifier scope preference (Morrill, 2000) ; finally, there is a wide-coverage syntactic and semantic parsing for French language, named Grail, that allows researchers to design and experiment with multimodal categorial grammars which includes AB grammars as well. (Moot, 2010). Grail is a modern, flexible and robust parser/automated theorem prover that has around 900 different category types for French language and works efficiently with the newspaper articles. Categorial grammar is extracted semi-automatically from the Paris 7 tree-bank and its semantic lexicon maps combinations of words, part-of-speech tags and formulas to Discourse Representation Structures (Kamp & Reyle, 1993). [2]

Our strategy for resolving incomplete utterances is based on using AB grammars as the basis of our syntactic formalism on the one hand, and adopting the notion of unification that exists in RG algorithm (learning Rigid AB grammars) (Buszkowski & Penn, 1990; Buszkowski, 1987) plus some dynamic programming technique on the other hand. This approach, to the best of our knowledge, is a new research technique that can efficiently deal with incomplete utterances with missing categories while benefiting from the good properties in categorial grammars. As we will see, the time complexity for the algorithm that we use to find one missing category with $n$ number of words is $O(n^4)$. This research can be extended to more missing categories, and other kind of incomplete utterances as we will discuss in the conclusion section.

Since this paper is self-contained, we have introduced some basic notions and definitions in section (2) which is essential for understanding AB grammars and notion of categorial unification. Moreover,

1. See (Leacock *et al.*, 2010, Chapter 1) for detailed descriptions.

2. As for the behaviour of the Grail parser on the sentences containing an unknown word we should say that the wide-coverage version of the Grail parser uses a probabilistic model to propose one or more of the contextually most likely options. It is worth mentioning that our proposal can be used for finding missed words ; although it can be adopted for finding categories for unknown words as well.

There are a number of running examples and illustrations on other sections to make notions more clearer. Section (3) discusses unification technique of RG algorithm, and how this mechanism can be adopted for finding the missing category with some examples and the possible combinations of the structural tree. Section (4) illustrate informally how unification technique and the notion of dynamic programming can be tackled with using bottom-up parsing in a modified version of well-known dynamic CYK algorithm. Section (5) concludes our approach and explains possible direction in future studies. We have also included an appendix that contains the pseudo-code of our algorithm provided with discussions on the complexity of the algorithm.

## 2 Reminder of AB grammars and Unification

Here, AB grammars, and some essential notions are discussed very quickly. As highlighted, this formalism is very suitable for our purpose since it enjoys learn-ability from positive examples and it converges after learning from a fixed number of sentence structures. (Buszkowski & Penn, 1990; Kanazawa, 1998; Bonato, 2000)

**Definition 1 : Categories**

Taking Lambek's notation (Lambek, 1958), categories in AB grammars can be written as follows (Moot & Retoré, 2012, Chapter 1) :

$$L ::= P \mid (L\backslash L) \mid (L/L)$$

where P is the set of primitive category (such as np and S) while $B\backslash A$ and $A/B$ can be interpreted as the functors that take $B$ as their arguments on its left and right-hand side respectively ; in the all mentioned cases, the result would be $A$.

A function Lex is a lexicon which maps words to finite set of categories. An expression, that is a sequence of words or terminals $w_1, \cdots, w_n$ is of category $u$ whenever there exists for each $w_i$ a category $u_i$ in $Lex(w_i)$ such that $u_1 \cdots u_n \rightarrow u$ with the left or right elimination rule patterns, namely $(A/B)B \rightarrow A$ and $B(B\backslash A) \rightarrow A$.

It is worth mentioning that variable categories are also used such as $x_1, x_2, \cdots, x_n$ (or $y_1, y_2, \cdots, y_n$) in addition to above fixed category–or what can be properly named fixed learned categories with syntactic labels– for our future purposes in unification phase. See following example a lexicon and the related process which ends up to $S$ :

Example 1 : Lexicon

| | | |
|---|---|---|
| Jean : $np$ | aime : $(np\backslash S)/np$ | Marie : $np$ |

Jean aime Marie $\Longrightarrow np,\ (np\backslash S)/np,\ np \Longrightarrow np,\ (np\backslash S) \Longrightarrow S$

**Definition 2 : Structural trees with variable categories**

It is basically a tree-structured proof representation with variable categories instead of having fixed categories. The only exception is fixed category $S$ in the root of the tree which is naturally expected. Arguments will have unique variable while the functors would get the argument category as its sub-formula in the right or left side. All the possible structural trees of a sequence with 3 words are

represented in figure 1. For convenient, when we use the term structural tree we mean the structural tree with variable categories.

It can be easily noticed that the number of possible trees for a sequence of string with $n$ length would be $catalan(n-1) * 2^{n-1}$ in which $catalan(n-1)$ is the number of possible binary trees and $2^{n-1}$ is the number of possible operations for $\backslash$ and $/$ over each binary sub-trees. [3] In the case of a sentence with 3 unspecific words, the number of all possible structural trees is 8 ($= catalan(3-1) * 2^2$) as it can be observed in the figure 1.

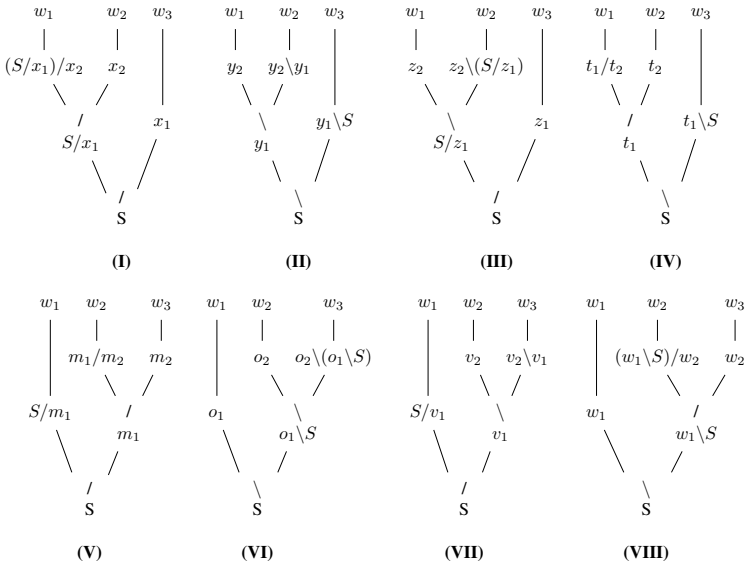



FIGURE 1 – Possible structural trees of a sequence $w_1 w_2 w_3$

| Words | Fixed learned categories | I | II | III | IV | V | VI | VII | VIII |
|---|---|---|---|---|---|---|---|---|---|
| Jean | $np$ | $(S/x_1)/x_2$ | $y_2$ | $z_2$ | $t_1/t_2$ | $S/m_1$ | $o_1$ | $S/v_1$ | $w_1$ |
| aime | $(np\backslash S)/np$ | $x_2$ | $y_2\backslash y_1$ | $z_2\backslash(S/z_1)$ | $t_2$ | $m_1/m_2$ | $o_2$ | $v_2$ | $(w_1\backslash S)/w_2$ |
| Marie | $np$ | $x_1$ | $y_1\backslash S$ | $z_1$ | $t_1\backslash S$ | $m_2$ | $o_2\backslash(o_1\backslash S)$ | $v_2\backslash v_1$ | $w_2$ |

TABLE 1 – Unification of variable categories and fixed (learned) categories

**Definition 3 : Pattern matching**

Let $t$ be a category which consists of at least one variable category and possibly some fixed ones, and let $L$ be a set of fixed categories, we say $t$ is matchable with set $L$, if there exist a substitution of $t$ which is a category in $L$ or a sub-category in $L$. For instance, if we define $L = \{np, S, S/(np\backslash S)\}$

3. The Catalan(n) is given in terms of binomial coefficients by following formula :

$$\mathrm{C}_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!\,n!} = \prod_{k=2}^{n} \frac{n+k}{k} \qquad \text{for } n \geq 0.$$

and $t = np\backslash x_1$, then, since the substitution $x_1 := S$ exists and $np\backslash S$ is a sub-category of $S/(np\backslash S)$ in $L$, we can say that $t$ is matchable with $L$ [4].

**Definition 4 : Derivation of variable categories**

Let $X$, $Y$ and $Z$ be categories with at least one primitive variable and $A$ a category with no primitive variable category (=fixed category). We can define, the derived result of $X$ and $A$ as $Y[X := (Y/A)]$; and the derived result of $A$ and $X$ as $Y[X := A\backslash Y]$; and finally the derived results of $X$ and $Y$ as $Z[Y := (X\backslash Z)]$ or $Z[X := (Z/Y)]$. The notation $t := t^{'}$ means substitution of $t$ with $t^{'}$.

Table 1 illustrates a fixed learned categories (the second column from left), and eight columns showing the variable categories (labeled from I to VIII) which are the yields of derivation trees in figure 1. We want to find one column, let say $n$ (between I to VIII), such that every line $k$ of the fixed learned categories unifies with the $k$ line of the $n$ column. It can be clearly observed that there is only one column (among eight possible cases) which is unifiable with the fixed learned categories and it is (VIII). By simple substitution we can observe that $w_1 = np$, $w_2 = np$, and since $(w_1\backslash S)/w_2$ matches with $(np\backslash S)/np$. One can observe how the rest seven cases fail in the unification phase. So, we can claim that structural tree (VIII) is unifiable with the lexicon and more important, the substitutions of $w_1 = np$ and $w_2 = np$ in the structural tree (VIII) yields its final derived tree.

It is worth mentioning that the technique described above can distinguish the same word used with different categories, and we will also see how the technique we propose can potentially be generalized to $k$ valued AB grammars in the conclusion section. Having these basic notations will allow us to start describing our techniques for dealing with uncompleted utterances.

# 3 Unification Technique of RG Algorithm

The unification technique which is used here is introduced for the first time in learning rigid AB grammars algorithm (Buszkowski & Penn, 1990) and it has been re-studied and extended by some researchers as well (Kanazawa, 1998; Bonato, 2000; Bonato & Retoré, 2014). We are using almost the same technique, but with a new purpose which is dealing with incomplete utterances with one missing category that to our knowledge is a new field to experience. Since we have explained the essential notions, we just focus on some examples to highlight the mechanism of unification in this section.

Let us consider the following lexicon which is adopted for our illustration purpose, and it is assumed that the lexicon is already learned from the positive sentences in corpus [5]. Now, we consider the following examples where a category is missing. One missing determiner in example (3) and one missing preposition in example (4) :

**Example 2 :** Lexicon

| | | | | |
|---|---|---|---|---|
| le , la : $np/n$ | dans : $(s\backslash s)/np$ | poisson , mer : $n$ | nage : $np\backslash S$ | vite : $(np\backslash S)\backslash(np\backslash S)$ |

**Example 3 :** * Poisson nage vite.

4. As might be noted, pattern matching here is very similar to unification. We have adopted this notation to make distinction between unification of two categories, and one category with some external references.

5. See (Moot & Retoré, 2012, Section 1.6.3) for technical details of learning AB rigid grammar; although, this is not essential for grasping the content of this section.

**Example 4 :** * Le poisson nage vite la mer.

Let us start with example (3). As it is illustrated in figure 2 we have illustrated only three cases. (Just note that without knowing the categories of words there would be 160 ($= catalan(3) * 2^3 * (3+1)$) possible functor-argument tree structures that can be associated to example 3).

We can unify the gained variable categories associated to the words in our structural trees with the one we have in our lexicon (figure 2).

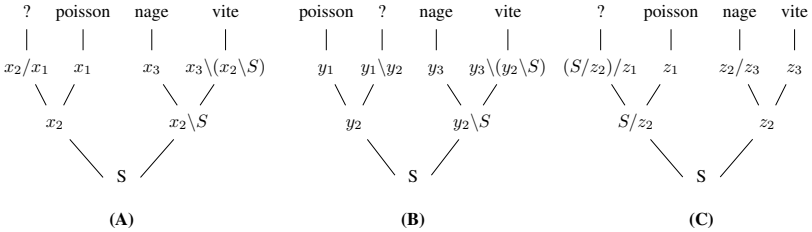



FIGURE 2 – Three possible structural trees for example 3

| Words | Learned categories | A | B | C |
|---|---|---|---|---|
| poisson | $n$ | $x_1$ | $y_1$ | $z_1$ |
| nage | $np\backslash S$ | $x_3$ | $y_3$ | $z_2/z_3$ |
| vite | $(np\backslash S)\backslash(np\backslash S)$ | $x_3\backslash(x_2\backslash S)$ | $y_3\backslash(y_2\backslash S)$ | $z_3$ |
| ? | – | $x_2/x_1$ | $y_1\backslash y_2$ | $(S/z_2)/z_1$ |

TABLE 2 – Unification of variable types and learned types

We can observe that structural tree (A) is unifiable and the solutions are as follows : $x_1 = n$, $x_2 = np$ and $x_3 = np\backslash S$ ; while the missed category can be gained by substitution of all occurrences of $x_i$ in variable category $x_2/x_1$ by its relevant fixed category which yields $np/n$ in our case. The suggested category with unification technique on this tree can only be licensed if it exists in our lexicon (or some valid external lexicon) which is the case. Notice that $np/n$ corresponds to determiner category.

Structural tree (B) is unifiable as well, and the unification technique yields $n\backslash np$ as a potential solution for the missing category with the same analysis. Since this solution does not exist in our lexicon it can not be accepted. Structural tree (C) is not unifiable and it is rejected without further analysis. We can analyze example (4) with the same technique. The problem is large number of its possible combinations of its structural tree. We will see a treatment for this issue in the next section.

# 4 AB grammars, Unification and Dynamic Programming

As we saw in the previous section, for (unknown) $n$ sequence of words the possible structural trees are ($= catalan(n-1) * 2^{n-1}$) and also the number of positions for a missing category is $n+1$.[6] The final number of different combinations of a $n$ sequence of words with one missing category is

6. For $n$ number of words with one missing category, we will have $n-1$ potential solutions between the words, one in the beginning and another one in the end.

$= catalan(n) * 2^n * (n+1)$ which grows faster than exponential rate (UTH, 1973). Actually, we do not need to explore all the possible structural trees since we know that some of our categories are fixed. In this section, we illustrate how the problem in example (4) can be tackled with dynamic bottom-up parsing.

The idea behind our solution is to try all the (n+1) positions for the missing category and avoiding searching all the possible structural trees by early pattern matching and unification technique of RG algorithm using a modified version of CYK dynamic programming technique [7]. With this strategy we can save each step results for making decision as we go further. As we will see, this leads us to an efficient computational complexity of $O(n^4)$ with n number of words. The table 3 is one of the possible cases (out of eight) for parsing example (4) assuming that $x_1$ (=missing category) is between 'vite' and 'la'. We start parsing step-by-step to illustrate how it works. [8]

| **1** | $np/n$ | $n$ | $np\backslash S$ | $(np\backslash S)\backslash(np\backslash S)$ | $x_1$ | $np/n$ | $n$ |
|---|---|---|---|---|---|---|---|
| | **Le** | **poisson** | **nage** | **vite** | **?** | **la** | **mer** |
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |

TABLE 3 – Step (1) of parsing with unification for example (4)

The table 3 shows the initial phase of parsing, in which the fixed rigid learned categories are assigned plus $x_1$, a variable category, which is supposed to be specified. [9]

| **2** | $np$ | – | $np\backslash S$ | $?\, x_2[x_1 := ((np\backslash S)\backslash(np\backslash S))\backslash x_2]$ | $?\, x_3[x_1 := x_3/(np/n)]$ | $np$ | |
|---|---|---|---|---|---|---|---|
| **1** | $np/n$ | $n$ | $np\backslash S$ | $(np\backslash S)\backslash(np\backslash S)$ | $x_1$ | $np/n$ | $n$ |
| | **Le** | **poisson** | **nage** | **vite** | **?** | **la** | **mer** |
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |

TABLE 4 – Step (2) of parsing with unification for example (4)

The table 4 shows step 2, i.e. all possible derivations of two sequential categories. Since, we have the variable categories, we can have the flexibility of working with unification technique. For example, variables $x_2$ and $x_3$ would be a possible derivation if $x_1$ be substituted with $x_3/(np/n)$ and $((np\backslash S)\backslash(np\backslash S))\backslash x_2$ respectively.

Notice that we have included question mark(=?) in the cells of $x_2$ and $x_3$ to signify that we have not performed the pattern matching yet. Our algorithm benefits from early pattern matching of variable categories with the lexicon or any external reference that provides categories [10]. Note that all the possible substitutions for $x_1$, namely $((np\backslash S)\backslash(np\backslash S))\backslash x_2$ and $x_3/(np/n)$ fail in pattern matching phase; since, there is no fixed type category (and even none subcategory) in our lexicon that can be matched with them. Early matching is an efficient technique since it prevents unnecessary computation. As depicted in step (3) in table 5 all possibilities for $x_1$ (in row 2) are ruled out.

7. See appendix for the pseudo-code of the algorithm

8. We have used bottom-up representation which is not very intuitive. This choice is made since it is in alignment with chart representation in dynamic programming.

9. Note that, in principle, we should assume all the possible positions of the unknown missing word which is 7 in this case. We have shown only one of the cases that leads to a solution. Although, our proposed algorithm would consider all the cases.

10. Until now, Moot's Grail (Moot, 2010) provides around 900 categories for French language. This can be used for pattern-matching in this phase as the external reference.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **3** | $S$ | – | $?\, x_3[x_1 := (np\backslash S)/x_3]$ | – | $x_2[x_1 := x_2/np]$ | | |
| **2** | $np$ | – | $np\backslash S$ | – | – | $np$ | |
| **1** | $np/n$ | $n$ | $np\backslash S$ | $(np\backslash S)\backslash(np\backslash S)$ | $x_1$ | $np/n$ | $n$ |
| | **Le** | **poisson** | **nage** | **vite** | **?** | **la** | **mer** |
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |

TABLE 5 – Step (3) of parsing with unification for example (4)

Again, we will follow the same procedure that we practiced for step 3 in table 5. In this step, all the possible derivations of three possible sequential categories are shown in each cell. The variable category $(np\backslash S)/x_5$ does not match with any fixed category in our lexicon, so, it will be ruled out. Although, $x_6/np$ matches with some categories in our lexicon so we will keep it as we go to the next step. We will continue the same procedure until we will reach to step 7 which is the last one. Table 6 shows the last step.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **7** | $x_5[x_3 := x_5/np]$<br>$x_4[x_2 := S\backslash x_4]$ | | | | | | |
| **6** | – | – | | | | | |
| **5** | $x_3[x_1 := S\backslash x_3]$ | – | – | | | | |
| **4** | $S$ | – | – | – | | | |
| **3** | $S$ | – | – | – | $x_2[x_1 := x_2/np]$ | | |
| **2** | $np$ | – | $np\backslash S$ | – | – | $np$ | |
| **1** | $np/n$ | $n$ | $np\backslash S$ | $(np\backslash S)\backslash(np\backslash S)$ | $x_1$ | $np/n$ | $n$ |
| | **Le** | **poisson** | **nage** | **vite** | **?** | **la** | **mer** |
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |

TABLE 6 – Last step of parsing with unification for example (4)

In table 6, all the early pattern matching is finished, and we should perform the unification. We can find two following results :

— In row 7, what we expect to see is $S$, so, $x_5 = S$. By substitution we have $x_3 = S/np$, so, $x_1 = S\backslash(S/np)$, since, this category does not exist in our lexicon it is ruled out.
— The same as above, in row 7, what we expect to see is $S$. By substitution we have $x_4 = S\backslash S$, so, $x_1 = (S\backslash S)/np$, since, this category exist in our lexicon it accepted as a potential answer, and this is what we were initially looking for, a determiner.

In this section, we informally analyzed how our modification to CYK algorithm can be useful in finding a missing category using early pattern matching and unification in learning RG algorithm. The relevant algorithm and its computational complexity are discussed in the appendix.

# 5 Conclusion and Possible Extensions

We explained how we can find one missing category in an incomplete utterance. The method is basically designed for AREN project (ARgumentation Et Numérique) in which the average number of words in a sentence is twenty and the number of categorial types are around two hundred, so our proposal efficiently works on this specific domain. There are four possible extensions for our future

study and research.

The first extension is moving to non-rigid categories, namely fixed number of several categories assigned to each single word– instead of rigid ones. Since our approach benefits from chart parsing, this is, in principle, tractable. However, it will increase our computational complexity as it is expected. We can reduce our search space when we have lexical ambiguities (several categories for each word of the input sentence) by adopting super-tagging technique (Bangalore & Joshi, 2011). The one which is very suitable for our purpose is Moot's super-tagger (Moot, 2015) which is implemented in the wide-coverage version of the Grail parser.

The second extension is to focus on several numbers of missing categories, instead of just one. Of course, we need to limit the number of missing categories, since, comprehension of sentences with so many missing categories is so hard even for human users of language, nevertheless, it will be useful if we stay limited to fixed number of categories. One, primitive suggestion would be trying all the possible combinations of $k$ missing words. This solution would have some drawbacks since we need to know the accurate number of missing categories, and more important, we should consider the increasing number of categories. It is difficult to rule out these categories by early pattern matching. One of the approaches would be using Jenson's approach (Jensen *et al.*, 1983) to build possible structural sub-trees of an incomplete utterance to find a limited number of suggestion for the position of missing categories. Further study is required for more clarification.

The third extension is to tackle other classes of incomplete utterances such as extra categories, misused categories, swap categories or any possible combinations. The main problem with this approach is the need of having a computational mechanism to know what is the exact problem of a sentence before solving it. For instance, if we know that in a given utterance, we have the fixed numbers of missing words, extra words, and swap categories, then we can tackle the problem more efficiently with a different module that we might build. The author has designed an algorithm that transforms a Context Free Grammar into Constraint Handling Rules (implemented in Prolog) for a fragment of natural language with relative clauses. Nevertheless, still, we need to have more research and study to analyses the efficiency and suitability of this kind of approaches comparing to the approaches that do not consider such distinctions.

The fourth extension is working on other families of categorial grammars such as Lambek's Calculus (Lambek, 1958) or some restricted version of Partial Proof Trees (Joshi & Kulick, 1997) for categorial grammar. The main problem –that the author is still dealing with– is that partial proof trees need to be restricted in order to work with Lambek's calculus. Moreover, designing a computationally decidable system with efficient complexity for incomplete utterances is still an on-going research. Generally speaking, this line of research can possibly be helpful in the domain of incoherence discourses. For instance, writing-assistant systems with questionnaires, for some specific targeted users such as patients suffering from specific mental disorders.

As discussed in the introduction, categorial grammar is chosen for its excellent correspondence with semantic. We drew the attention of readers to a special case, namely missing quantifier in the incomplete utterances. Once the missing quantifier is found out, we can make a complete sentence, and deduce different semantic readings of the utterance. Since each semantic proof corresponds to a cut-free proof net, we can make different complexity profiles having different proof nets gained from different readings. The low complexity profile will have the preferred reading among other readings(Morrill, 2000). This is the main role of categorial grammars in modeling preferences of semantic readings that naturally makes our syntax-semantic interface work.

# 6 Acknowledgements

I would like to thank Christian Retoré for inviting me to his lecture on learning algorithms for categorial grammars at Montpellier University and also for his advice on the possibility of using the learning algorithms for addressing the problem of sentence completion.

# 7 Appendix : Algorithm

The proposed technique for finding a missing category is not implemented yet. However, algorithm ( 1) written in high-level pseudo-code is provided. The underlying technique namely the unification is illustrated in details with running examples in the previous sections. One can observe that the complexity is basically similar to CYK standard algorithm which is $O(n^3)$ with the exception that it runs one extra loop with $n$ steps, which makes the complexity of our algorithm $O(n^4)$ in the worst case scenario.

**input :** An array 'words' and an array 'lexlist' of the categories
**output :** An array [(pos,cat)] of the pairs indicating position and categories of the missing words

```
result ← [ ];
n ← length(words);
for t ← 1 to n + 1 do
    k ← 1;
    ylist ← null;
    n_words = insert(words, x[k], t);
    // initializes the first row;
    for r ← 1 to n + 1 do
        p [1, r, 1] = lex(n_words[r], null, null)
    end
    for i ← 2 to n + 1 do
        for j ← 1 to n − i + 2 do
            for m ← 1 to i − 1 do
                // performs early pattern matching and unification;
                xlist = unify(p[m, j, 1], p[i − m, j + m, 1], lexlist, k) ;
                if xlist != fail then
                    ylist = ylist + xlist ;
                    assign(p[i, j], xlist) ;
                    else p[i,j,1]=null ;
                end
            end
        end
    end
    // substitutes the variables with its content;
    for v ← k to 1 step −1 do
        t2 ← snd(ylist(v));
        t3 ← trd(ylist(v));
        if is_fixed(t3) then
            substitute_all(ylist, t2, t3)
        end
    end
    // selects all the appropriate candidates for missing categories;
    for w ← 1 to len(ylist) do
        t2 ← snd(ylist(v));
        t3 ← trd(ylist(v));
        if (is_fixed(t3) and t2=x[1] and is_in_lexicon(lexlist,t3)) then
            result = result + [(t, t3)]
        end
    end
    return (result)
end
```

**Algorithm 1:** An algorithm for the new proposal

# Références


BANGALORE S. & JOSHI A. (2011). *Supertagging : Using Complex Lexical Descriptions in Natural Language Processing*. MIT Press.

BLACHE P. (2000). Constraints, linguistic theories, and natural language processing. In *International Conference on Natural Language Processing*, p. 221–232 : Springer.

BONATO R. (2000). *Uno studio sull'apprendibilità delle grammatiche di Lambek rigide—a study on learnability for rigid Lambek grammars*. PhD thesis, Tesi di Laurea & Mémoire de DEA, Università di Verona & Université Rennes 1.

BONATO R. & RETORÉ C. (2014). Learning lambek grammars from proof frames. In *Categories and Types in Logic, Language, and Physics*, p. 108–135. Springer.

BUSZKOWSKI W. (1987). Discovery procedures for categorial grammars. *Categories, Polymorphism and Unification*, p. 35–64.

BUSZKOWSKI W. & PENN G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica*, **49**(4), 431–454.

DAHL V. & BLACHE P. (2004). Directly executable constraint based grammars. In *Proc. Journees Francophones de Programmation en Logique avec Contraintes, Angers, France*, p. 149–166.

DINI L. & MALNATI G. (1993). Weak constraints and preference rules. *Studies in Machine Translation and Natural Language Processing*, p. 75–90.

JENSEN K., HEIDORN G. E., MILLER L. A. & RAVIN Y. (1983). Parse fitting and prose fixing : getting a hold on ill-formedness. *Computational Linguistics*, **9**(3-4), 147–160.

JOSHI A. K. & KULICK S. (1997). Partial proof trees as building blocks for a categorial grammar. *Linguistics and Philosophy*, **20**(6), 637–667.

KAMP H. & REYLE U. (1993). From discourse to logic ; introduction to the modeltheoretic semantics of natural language.

KANAZAWA M. (1998). *Learnable Classes of Categorial Grammars.* ERIC.

LAMBEK J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly*, **65**(3), 154–170.

LEACOCK C., CHODOROW M., GAMON M. & TETREAULT J. (2010). Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, **3**(1), 1–134.

MELLISH C. S. (1989). Some chart-based techniques for parsing ill-formed input. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, p. 102–109 : Association for Computational Linguistics.

MOOT R. (2010). Wide-coverage French syntax and semantics using Grail. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal. System Demo.

MOOT R. (2015). A type-logical treebank for french. *Journal of Language Modelling*, **3**(1), 229–264.

MOOT R. & RETORÉ C. (2012). *The logic of categorial grammars : a deductive account of natural language syntax and semantics*, volume 6850. Springer.

MORRILL G. (2000). Incremental processing and acceptability. *Computational linguistics*, **26**(3), 319–338.

PROST J.-P. (2009). *Modelling Syntactic Gradience with Loose Constraint-based Parsing 8 Modélisation de la gradience syntaxique par analyse relâchéea base de contraintes*. PhD thesis, Citeseer.

PULLUM G. K. & SCHOLZ B. C. (2001). On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In *International Conference on Logical Aspects of Computational Linguistics*, p. 17–43 : Springer.

SCHNEIDER D. & MCCOY K. F. (1998). Recognizing syntactic errors in the writing of second language learners. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, p. 1198–1204 : Association for Computational Linguistics.

UTH D. (1973). The art of computer programming, vol 1 : Fundamental algorithms.

VOGEL C. & COOPER R. (1995). Robust chart parsing with mildly inconsistent feature structures. *Nonclassical feature systems*, **10**, 197–216.

# Expressions polylexicales verbales : étude de la variabilité en corpus

Caroline PASQUER[1]
(1) LI - Laboratoire d'Informatique de l'Université de Tours
BDTLN - Bases de données et traitement des langues naturelles
caroline.pasquer@etu.univ-tours.fr

RÉSUMÉ

La reconnaissance et le traitement approprié des expressions polylexicales (EP) constituent un enjeu pour différentes applications en traitement automatique des langues. Ces expressions sont susceptibles d'apparaître sous d'autres formes que leur forme canonique, d'où l'intérêt d'étudier leur profil de variabilité. Dans cet article, nous proposons de donner un aperçu de motifs de variation syntaxiques et/ou morphologiques d'après un corpus de 4441 expressions polylexicales verbales (EPV) annotées manuellement. L'objectif poursuivi est de générer automatiquement les différentes variantes pour améliorer la performance des techniques de traitement automatique des EPV.

ABSTRACT

**Verbal MWEs : a corpus-based study of variability**

Identification and processing of multiword expressions (MWEs) are required by many natural language processing applications. MWEs can appear in other forms than their canonical form, hence the interest of studying their variability profile. In this paper, we present an overview of syntactic and / or morphological variation patterns based on a corpus of 4441 manually annotated verbal MWEs. Our goal is to automatically generate variants to improve the performance of MWE automatic processing techniques.

MOTS-CLÉS : Expressions polylexicales, figement, variabilité.

KEYWORDS: Multiword expressions, MWE, variability.

## 1 Introduction

Les expressions polylexicales (EP ou *multiword expressions, MWE* en anglais) sont des groupements de mots dont le sens individuel "ne permet pas d'interpréter l[a] combinaison" (Gross, 1982) comme dans l'expression *casser du sucre sur le dos* où il n'est pas plus question de *sucre* que de *dos*. Cette spécificité (nommée *non-compositionnalité*) pose problème pour différentes applications de traitement automatique des langues (par exemple l'impossibilité de traduction automatique "mot à mot" : *#to break sugar on one's back* [1]).

La détection des EP est une question d'autant plus importante qu'elles sont très fréquentes – 40 % des mots du journal *Le Monde* font partie d'EP (Gross & Senellart, 1998). Jackendoff (1997) estime d'ailleurs que chaque langue contient autant d'EP que de mots isolés. De plus, il est impossible

1. Nous adoptons la convention suivante : le symbole # désigne un glissement de sens : *prendre congé (= dire au revoir)* versus # *prendre des congés (= des vacances)* ; l'astérisque * indique une agrammaticalité (**il faut que je viens*).

d'en établir une liste exhaustive car de nouvelles expressions apparaissent sans cesse, qu'il s'agisse d'entités nommées polylexicales (noms de personnes, d'organisations, etc.) ou de néologismes obtenus en réutilisant le lexique disponible (*canon à électron* cité par Gross (1982)). Baldwin *et al.* (2004) indiquent qu'un défaut de détection des EP est responsable de 8 % d'erreurs de parsing. La prise en compte du caractère évolutif des EP justifie les nombreux travaux consacrés à leur détection automatique.

On peut supposer que, si la variabilité des EP est qualitativement et quantitativement hétérogène, cette variabilité est plus élevée pour les EP verbales (EPV) dont les emplois peuvent différer de leur forme canonique (répertoriée dans un lexique par exemple) en raison de discontinuités, passivation, etc. : ***adresser*** *des* ***reproches*** → *des* ***reproches*** *semblables ont d'abord été* ***adressés*** *à certaines sectes*.

Le système le plus performant en matière de détection d'EPV parmi les sept testés dans le projet PARSEME (décrit en section 4) ne reconnaît que 55 % de variantes d'EPV déjà vues durant l'entraînement, soit presque 2 fois moins que pour des formes de surface identiques (90 % détectées). On émet l'hypothèse que la description des motifs de variabilité des EPV permettra le rapprochement de variantes quelque soit cette forme de surface. De plus, la connaissance des variantes possibles mettra en lumière les accords interdits dans les EPV afin de différencier une lecture littérale d'une lecture idiomatique (*je* ***vide*** *mon* ***sac/#ton sac***).

Ainsi, cet article a pour objet l'analyse et la caractérisation de la variabilité (morphologique et/ou syntaxique) d'EPV attestées en corpus. Cette analyse pourra être mise à profit pour différentes applications : génération de variantes à partir de la forme canonique ou, à l'inverse, fusion de variantes avec leur forme canonique. Ces deux tâches permettront d'améliorer la couverture des systèmes de détection d'EP. D'autres applications sont possibles telles que l'annotation automatique d'EP dans des corpus arborés ou dans des sorties de parseurs, l'*entity linking* ou la résolution de coréférences. La mise en évidence de spécificités de certains types d'EPV (constructions à verbe support par exemple) pourrait aussi permettre un traitement sémantique adéquat.

Nous revenons (section 2) sur la définition des EP et faisons un état de l'art sur leur variabilité (section 3). Le corpus est présenté en section 4 et la méthodologie d'exploitation des données en section 5. Nous détaillons la variabilité interne (morphologique) des EPV en section 6 et la variabilité externe (insertions) en section 7. Enfin, nous nous intéressons en section 8 aux dépendances syntaxiques dans des expressions de type verbe-(déterminant)-nom (abrégées sous la forme VB-(DET)-NOM) pour en extraire des profils de variation.

## 2 Définition et classification des EP

Malgré les controverses liées à la définition des EP (Villavicencio *et al.*, 2007; Savary, 2008), Savary (2008) observe que les linguistes s'accordent pour qualifier d'EP les expressions satisfaisant les 3 critères suivants :

— le fait de contenir au moins 2 unités lexicales ("mots"),
— dénotation unique et constante,
— idiosyncrasie morphologique, syntaxique, distributionnelle ou sémantique.

Sag *et al.* (2002) distinguent ainsi les **expressions institutionnalisées** des **phrases lexicalisées**. Les phrases institutionnalisées (ou collocations) ne satisfont pas le critère de non-compositionnalité et se caractérisent par la tendance des locuteurs à associer certains mots de façon préférentielle (par exemple : *aimer à la folie*).

Sag *et al.* (2002) analysent la variabilité (nommée *flexibilité* dans leur terminologie) des phrases lexicalisées – autrement dit des EP – en les classant par ordre de flexibilité croissante :

— Les **expressions figées** (par exemple *revenons à nos moutons*) n'admettent aucune variation morphosyntaxique : impossibilité d'insertion ou de flexion (#*revenez encore à mon mouton*).

— **Expressions semi-figées** : ces expressions présentent des contraintes strictes quant à l'ordre des mots ou leur composition mais autorisent certaines variations telles que le choix du déterminant ou la flexion temporelle : ***je tire / tirerai mon épingle du jeu***. Sag *et al.* (2002) utilisent le principe de compositionnalité sémantique défini par Nunberg *et al.* (1994) – le sens global d'une expression est lié à celui des éléments qui la composent – pour distinguer les idiomes sémantiquement décomposables et non-décomposables. Les *idiomes non-décomposables* (Sag *et al.*, 2002) font partie des expressions semi-figées et se caractérisent par leur opacité sémantique. Cette opacité impliquerait une variabilité limitée, par exemple une impossibilité de passivation : *#son épingle a été tirée du jeu*.

— **Expressions syntaxiquement flexibles** : figurent notamment ici les constructions à verbe support et les idiomes décomposables par équivalence sémantique (***spill the beans*** → *spill = reveal ; beans = secret(s)*). La distinction idiomes décomposables/non-décomposables est cependant remise en question (Abeillé, 1988; Sheinfux *et al.*, 2017) en raison de contre-exemples d'idiomes non-décomposables et pourtant flexibles.

Malgré ces contre-exemples, l'intérêt de la classification de Sag *et al.* (2002) est de considérer la flexibilité (ou variabilité) comme un critère essentiel de classification des EP. Nous adoptons la terminologie utilisée lors de la campagne d'annotation [2] du corpus PARSEME pour 18 langues (Savary *et al.*, 2017), selon laquelle les constituants requis et non substituables d'une EP sont qualifiés d'éléments *lexicalisés*. Ils seront signalés en caractères italiques gras dans les exemples. Nous adoptons également la typologie des EPV couvrant :

— Les constructions à verbe support, désormais abrégées **LVC**, de l'anglais *Light Verb Construction* : leur particularité réside dans le fait que ce n'est pas le verbe qui remplit la fonction de prédicat de la phrase, mais un nom prédicatif (***faire allusion***).

— Les idiomes, désormais abrégés **ID** : leur sens n'est pas compositionnel et ils possèdent souvent une double lecture littérale / idiomatique (***tourner la page***) nécessitant une désambiguïsation contextuelle. La classification des EPV dans le corpus relève d'un accord entre les 18 langues participant au projet et ne coïncide pas systématiquement avec celle traditionnellement admise (par exemple ***faire partie*** est un ID et non pas une LVC).

— Les verbes intrinsèquement réflexifs, désormais abrégés **IReflV**, de l'anglais *Inherently Reflexive Verbs*. Ils satisfont l'une des trois conditions suivantes : soit ils possèdent un sens différent du verbe seul (hors pronom), comme ***se rendre /*** *rendre*, soit ils n'existent pas sans *se* : ***se prélasser /*** **prélasser*, soit – tout en conservant le même sens – ils possèdent un cadre de sous-catégorisation spécifique (***se confesser*** *de X / confesser X*).

# 3 État de l'art sur la variabilité des EP

Quoique des travaux aient porté sur la question des EP dès les années 1980 (Gross, 1982), Sag *et al.* (2002) l'ont remise sur le devant de la scène avec leur classification des EP. Dès lors, les EP ont été examinées sous différents points de vue, qu'il s'agisse de la nature des EP (nominales, verbales, etc.) ou de certains types de variations : morphologiques (Nissim & Zaninello, 2013), syntaxiques ou sémantiques (Jacquemin, 2001). Ces différentes approches répondent à des objectifs différents :

2. Guide d'annotation PARSEME : `http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/`

— La **description lexicale** : Lichte *et al.* (2017) ont répertorié différents formalismes aptes à rendre compte des variantes des EPV. Le codage Walenty a par exemple permis de constituer le *Polish Valence Dictionary* comportant 1774 verbes. 8000 EPV y sont décrites avec leur cadre valenciel et leurs motifs de variabilité. Ce formalisme nécessite des ajustements pour couvrir la diversité des variantes possibles (Przepiórkowski *et al.*, 2014) surtout si on souhaite l'extrapoler à une langue différente du polonais pour lequel il a été initialement conçu (par exemple : absence de déclinaisons en français mais présence de déterminants).
— L'**analyse de variabilité en corpus** : Nissim & Zaninello (2013) se sont intéressés à la variation interne d'EP nominales en italien et proposent une méthode de recherche flexible grâce à des patrons de variation pour améliorer la détection des EP. Pour le français, Tutin (2016) a établi des niveaux de variabilité pour les 30 EPV (ID et LVC) les plus fréquentes de type VB-(DET)-NOM (détails en section 8).
— La **prise en compte de la variabilité pour les applications** : dans le cas d'EP nominales, l'outil FASTER (Savary & Jacquemin, 2003) permet la reconnaissance de termes appartenant à une liste pré-établie en détectant les variantes des termes en corpus et en extrayant les variations par des métarègles morpho-syntaxiques et syntaxico-sémantiques.

Notre travail s'inscrit dans la continuité de ces approches en visant une description en corpus de la variabilité des EPV.

# 4 Corpus

Le corpus utilisé provient de l'identification manuelle d'EPV effectuée dans le cadre du projet PARSEME [3] (*PARSing and Multi-word Expressions*). Un guide d'annotation des EPV a été élaboré pour 18 langues avec des arbres de décisions permettant d'identifier puis de classer les EP. Une description détaillée de l'annotation et du corpus PARSEME est donnée par Candito *et al.* (2017). Tous les exemples mentionnés dans cet article proviennent de ce corpus à l'exception des exemples 3, 4 et 5 indiqués comme provenant du Web.

Le corpus d'entraînement français de PARSEME comporte 17 880 phrases résultant de la fusion des corpus Séquoia (Candito & Seddah, 2012) et UD (Nivre *et al.*, 2016) – correspondant à 450 221 tokens – et contient 4 441 occurrences d'EPV uniques [4]. 2,2 % des tokens font partie d'EPV, ce qui laisse supposer que la majorité des EP ne sont pas verbales si l'on se réfère au pourcentage de 40 % d'EP relevé par Gross. Quoique de taille réduite, ce corpus est sans équivalent dans l'espace francophone. Il offre l'intérêt de fournir des informations détaillées établies manuellement, qu'il s'agisse d'étiquetage morphosyntaxique ou des relations de dépendances syntaxiques formulées avec les tagsets de *Universal Dependencies* [5] et de Sequoia [6]. Ces deux types d'étiquetages ont été harmonisés. Le chevauchement se produisant lorsqu'un même élément est commun à plusieurs EPV est pris en compte (dans la phrase *Les **demi-finales** et la **finale** se **jouent** à Copenhague*, 2 EPV sont annotées : ***jouer demi-finales*** et ***jouer finale***). Les trois types d'EPV (ID, LVC, IReflV) y sont présents dans des proportions relativement équilibrées (Table 1). L'unique mention **OTH** (OTHER) correspond à l'expression ***aller et venir***. Si l'on s'intéresse aux LVC et aux ID (en dehors des impersonnels : ***il faut, il s'agit, ...***), la majorité des patrons (72 %) est du type VB-(DET)-NOM.

3. http://typo.uni-konstanz.de/parseme/
4. Le corpus contenait initialement 4 462 occurrences mais 21 occurrences ont été supprimées en raison de la présence de quelques phrases en double.
5. http://universaldependencies.org
6. https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia

| Type d'EPV | ID | LVC | IReflV | OTH | TOTAL |
|---|---|---|---|---|---|
| nombre d'occurrences | 1777 occ. | 1357 occ. | 1306 occ. | 1 occ. | 4441 occ. |

TABLE 1 – Répartition des différents types d'EPV dans le corpus.

| ID | | | | LVC | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***avoir lieu*** | 95 occ. | ***faire partie*** | 88 occ. | ***jouer rôle*** | 33 occ. | ***faire appel*** | 32 occ. | ***prendre mesure*** | 11 occ. |
| ***faire l'objet*** | 27 occ. | ***mettre fin*** | 23 occ. | ***avoir besoin*** | 28 occ. | ***avoir droit*** | 17 occ. | ***marquer but*** | 10 occ. |
| ***faire face*** | 20 occ. | ***porter nom*** | 19 occ. | ***jouer match*** | 16 occ. | ***faire apparition*** | 15 occ. | ***avoir effet*** | 10 occ. |
| ***prendre part*** | 19 occ. | ***tenir compte*** | 17 occ. | ***signer contrat*** | 15 occ. | ***disputer match*** | 15 occ. | | |
| ***mettre un terme*** | 13 occ. | ***prendre nom*** | 13 occ. | ***poser question*** | 14 occ. | ***faire référence*** | 13 occ. | | |
| ***donner lieu*** | 12 occ. | ***donner naissance*** | 10 occ. | ***rendre hommage*** | 12 occ. | ***lancer appel*** | 11 occ. | | |
| ***avoir mal*** | 9 occ. | ***faire preuve*** | 9 occ. | ***avoir chance*** | 11 occ. | ***avoir tendance*** | 11 occ. | | |
| TOTAL ID : 374 occ. (21 % des ID du corpus) | | | | TOTAL LVC : 274 occ. (20 % des LVC du corpus) | | | | | |

TABLE 2 – Liste des EPV de type VB-(DET)-NOM les plus fréquentes en corpus.

La Table 2 en montre les 648 occurrences les plus fréquentes.

Si l'on s'intéresse aux éléments lexicalisés pour identifier des patrons d'EPV [7] :

- **IReflV** : 1299 occ. Après vérification manuelle de 3 cas non conformes en raison d'erreurs d'annotation, 100 % des IReflV sont de type PRON-VB ou VB-PRON (***souvenez-vous***).
- **LVC** : 1298 occ. (95 % des LVC) sont du type VB-NOM ou NOM-VB et 30 occ. (2 %) du type VB-PREP-(DET)-NOM, d'où notre étude principalement focalisée sur le patron VB-NOM/NOM-VB.
- **ID** : on observe davantage de variation dans les patrons (84 différents). Le plus fréquent ne représente que 31 % (549 occ.) : il s'agit de VB-NOM *(**signer** son **retour***). La catégorie VB-DET-NOM, où DET est lexicalisé, correspond à 13 % des occ. (233 occ.). Enfin, 180 occ. (10 %) sont du type VB-PREP-(DET)-NOM (***appeler à secours, revenir à la raison***).

Hormis pour les IReflV dont la structure en PRON-VB est prévisible, le patron ne permet pas de distinguer les ID des LVC : le schéma VB-(DET)-NOM par exemple figure dans ces deux catégories. De même, le verbe présent dans l'EPV ne suffit pas à lui seul à distinguer les ID des LVC puisque les verbes majoritairement utilisés (*avoir* et *faire*) y ont une distribution similaire : respectivement 19 % et 16 % des verbes de ces deux catégories.

# 5 Méthodologie

L'annotation manuelle des EPV met en évidence les éléments lexicalisés. Le regroupement des EPV en fonction des éléments lexicalisés – en neutralisant leur ordre d'apparition ***jouer** un **rôle** / le **rôle** est **joué*** – conduit à l'obtention de 1584 entrées distinctes. La limite de cette approche concerne la possibilité que des constituants identiques renvoient à des EPV différentes. Ce risque est cependant très limité puisque sur les 31 EPV étudiées en détail (Table 2), seule l'entrée (*faire ;appel*) illustre ce cas de figure :

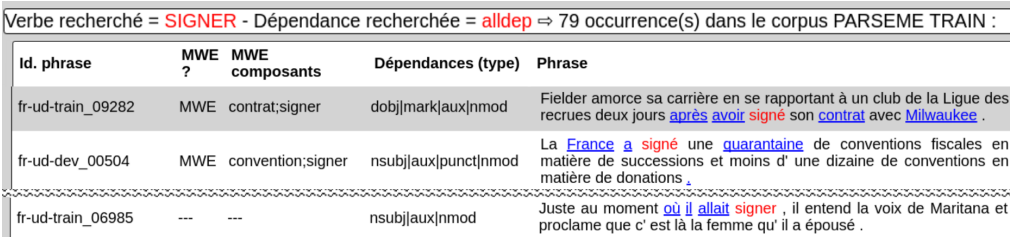
**Ex. 1.** *faire appel* = 'contester en justice' (4 occ.) → *Les deux pilotes déclassés **firent appel** Ø, appel jugé en janvier 2007.* (corpus PARSEME)

**Ex. 2.** *faire appel (à)* = 'recourir à' (28 occ.) → *Je vous déconseille vivement de **faire appel** à cette agence.* (PARSEME)

7. VB = verbe ; PRON = pronom ; DET = déterminant ; PREP = préposition ; ADJ = adjectif ; ADV = adverbe ; CONJ = conjonction ; AUX = auxiliaire ; REL = relatif

Si les éléments annotés mettent en évidence la structure syntaxique figée des EPV (section 6), les éléments non annotés (éléments insérés entre les éléments lexicalisés) sont également instructifs. En se restreignant aux EPV de type VB-(DET)-NOM (en raison de leur fréquence élevée), on obtient un premier aperçu de la variabilité reposant à la fois sur la fréquence et sur la nature des insertions linéaires (section 7).

Nous avons également mis à profit les dépendances syntaxiques fournies dans le corpus. Nous avons développé un système de requêtes (Figure 1) pour montrer les différents types de dépendances d'un nom ou d'un verbe lorsqu'il fait partie (ou non) d'une EPV. Pour cela, l'utilisateur spécifie le verbe ou le nom recherché et éventuellement un type de dépendances. Les phrases correspondantes s'affichent en indiquant s'il s'agit ou non d'une EPV. Dans chacun de ces cas, la nature et la fréquence des différents types de dépendances sont mentionnées.


Verbe recherché = SIGNER - Dépendance recherchée = alldep ⇨ 79 occurrence(s) dans le corpus PARSEME TRAIN :

| Id. phrase | MWE ? | MWE composants | Dépendances (type) | Phrase |
|---|---|---|---|---|
| fr-ud-train_09282 | MWE | contrat;signer | dobj\|mark\|aux\|nmod | Fielder amorce sa carrière en se rapportant à un club de la Ligue des recrues deux jours après avoir signé son contrat avec Milwaukee . |
| fr-ud-dev_00504 | MWE | convention;signer | nsubj\|aux\|punct\|nmod | La France a signé une quarantaine de conventions fiscales en matière de successions et moins d' une dizaine de conventions en matière de donations . |
| fr-ud-train_06985 | --- | --- | nsubj\|aux\|nmod | Juste au moment où il allait signer , il entend la voix de Maritana et proclame que c' est là la femme qu' il a épousé . |

FIGURE 1 – Capture d'écran du concordancier : recherche de toutes les dépendances (en bleu) du verbe SIGNER (en rouge). Le type de dépendances s'affiche dans une infobulle au passage de la souris.

# 6 Variation morphologique des constituants

Nous nous intéressons à la variation dite interne des EPV, c'est-à-dire à la capacité de flexion de ses constituants (nom ou verbe). L'étendue des flexions autorisées nous renseigne en effet sur le degré de variabilité d'une EPV (***porter ses fruits***/#*porter son fruit*).

## 6.1 Variabilité nominale

L'étude de la variation flexionnelle du nom a été restreinte aux EPV de type VB-(DET)-NOM les plus fréquentes (Table 2) en raison de leur forte représentativité dans le corpus. Cet examen porte sur 17 LVC (274 occ.) et 12 [8] des 14 ID (356 occ.). La taille du corpus est suffisante pour identifier des motifs de variabilité, mais le recours à des ressources complémentaires est nécessaire pour confirmer l'invariabilité. Ainsi, 12 LVC présentent une variabilité du nom dans notre corpus et 3 autres dans un moteur de recherche sur Internet (Ex. 3, 4, 5). Seules 2 LVC (***faire appel***, ***avoir*** *(le)* ***droit***) sont effectivement morphologiquement rigides.

**Ex. 3.** *Elle* ***fait*** *également une* ***apparition*** *au festival* (PARSEME) / *Cet artiste [...] a déjà* ***fait*** *des* ***apparitions*** *au Caveau de la Gare* (Web)

8. Le déterminant figé dans ***mettre un terme*** et ***faire l'objet*** signale une absence de flexion nominale.

**Ex. 4.** *l'occasion de lui* ***rendre*** *un dernier* ***hommage*** (PARSEME) **/** [...] *et leur* ***rend*** *les derniers* ***hommages***. (Web)

**Ex. 5.** ***J'ai*** *plutôt* ***tendance*** *à penser qu'il s'agit d'un coup monté.* (PARSEME) **/** ***J'ai*** *des* ***tendances*** *à être claustrophobe* (Web).

De la même façon, un examen des 11 ID apparemment figés met en évidence une flexibilité nominale de l'ID ***prendre nom***, d'où une variabilité nominale de 9,6 % des occurrences de type ID alors qu'elle est de 82 % pour les LVC.
Les contre-exemples mentionnés ne remettent pas en question la représentativité du corpus puisque 15 LVC (224 occ.) sur les 17 LVC et 11 ID (343 occ.) sur les 12 ID ont une flexibilité observée dans le corpus conforme à ce que l'on observe à plus grande échelle.

## 6.2 Variabilité verbale

Nous observons deux types de variabilité verbale dans le corpus :

— **Flexion personnelle et temporelle**
Parmi les cas d'invariabilité verbale, figurent des restrictions de flexion personnelle : verbes uniquement fléchis aux 3èmes personnes (*il(s)* ***se déroule(nt)/#je me déroule***), voire uniquement à la 3ème personne du singulier. Parmi les ID, on dénombre ainsi 518 (29 %) tournures impersonnelles (***il y a, il faut, il s'agit, il pleut des cordes, sera-t-il question***,...). Par ailleurs, certaines EPV ont une flexion temporelle limitée (***qu'importe/qu'importait/#qu'importa***) ou impossible (***honni soit qui mal y pense***).

— **Préfixation du verbe**
Le corpus comporte 7 EPV dans lesquelles le verbe est préfixé avec *r(e)-*, par exemple ***(re)donner raison*** ou ***(re)faire appel***. S'il s'agit bien ici de variantes, d'autres EPV nécessitent au contraire ce préfixe (***revenir à la raison / #venir à la raison***). La question de la double lecture littérale / EPV se pose également dans ***relever la tête*** puisqu'au sens littéral *lever* et *relever* sont des variantes alors que le préfixe est obligatoire dans l'EPV. Prêter attention à ce phénomène pourrait permettre, à partir d'un lexique préétabli d'EPV, d'en identifier des variantes pour des verbes a priori très productifs comme *faire* ou *donner*.

S'intéresser à la variabilité des EP ne se restreint pas à la variabilité de ses constituants : il faut aussi prendre en compte la variabilité externe, c'est-à-dire les discontinuités d'éléments lexicalisés qui constituent un problème majeur pour certains outils TAL (étiquetage séquentiel).

# 7 Discontinuités d'éléments lexicalisés

Les discontinuités d'éléments lexicalisés nous informent sur le degré de variabilité d'une EPV : plus elles sont nombreuses et variées, plus l'EPV est flexible. Par exemple, dans les phrases (6) et (7), toutes les deux du type *prendre* + NOM, la seconde tolère moins de variations :

**Ex. 6.** *il* ***prend*** *encore / son / un grand* ***essor*** (PARSEME : annotation alternant entre ID et LVC).

**Ex. 7.** *il* ***prend*** *encore / *sa / *une grande* ***conscience*** (PARSEME : ID).

Nous proposons dans ce qui suit de décrire les insertions linéaires observées afin d'en tirer des conclusions sur la variabilité des différents types d'EPV. On appellera fenêtre d'annotation, matérialisée

par les crochets (Ex. 8 et 9), l'ensemble des éléments compris entre le premier et le dernier élément annoté. Les éléments non annotés dans la fenêtre d'annotation sont qualifiés de discontinuités et sous-entendent une plage de variabilité. Dans ce qui suit, le nombre de discontinuités correspond au nombre total d'insertions, qu'elles soient contiguës (Ex. 8) ou non (Ex. 9) :

**Ex. 8.** *C'est ce que [**mettent** à mon avis **en lumière**] quelques propositions d'amendement.*(PARSEME)

**Ex. 9.** *Le président du tribunal [**procède** alors **à** la **lecture**] des chefs d'accusation.* (PARSEME)

Sur 4441 EPV, 2772 sont continues (62 %), et 1097 (25 %) ont une discontinuité d'un seul élément. Autrement dit, l'absence de discontinuités et l'insertion d'un seul élément couvrent à elles-seules 87 % du corpus. La Figure 2 montre bien que les ID sont légèrement moins figés de ce point de vue que les IReflV, mais bien davantage que les LVC. En effet, sur l'ensemble des LVC, seules 281 occ. (20,7 %) sont continues.

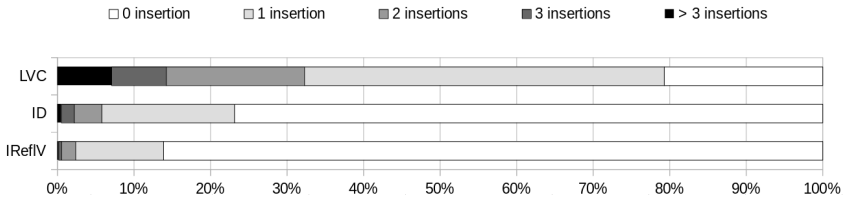



FIGURE 2 – Distribution des discontinuités dans les EPV du corpus.

Dans le cas des insertions uniques (Table 3), les différences sont significatives :

- 79 % de verbe *être* pour les IReflV. Il s'agit d'une variation aspectuelle régulière (par exemple *l'Empire [**s'est écroulé**]*),
- 66 % d'articles pour les LVC,
- 44 % d'adverbes et 28 % d'articles pour les ID. Cela rejoint le constat de Gross (1988) concernant la possibilité d'insertion d'adverbes dans les phrases figées (en dehors de quelques exceptions : *j'**aime** *vraiment **autant** vous le dire*). Quoique l'insertion d'un adverbe puisse être considérée comme une variation régulière (Tutin, 2016), la proportion d'adverbes dans les ID y est proportionnellement dix fois plus élevée que dans les LVC. Comme le déterminant est moins souvent lexicalisé dans les LVC, l'adverbe y est souvent associé. C'est ce que l'on observe pour les insertions de 2 éléments : sur 72 adverbes insérés, 4 insertions sont de type ADV-ADV et 42 de type ADV-DET. Par ailleurs, l'insertion moins fréquente de déterminant dans les ID est une propriété révélatrice car "l'extraction d'un nom est généralement interdite lorsque le déterminant est figé, et permise lorsque le déterminant et les modifieurs sont libres" (Laporte, 1988).

Un examen des éléments annotés met en évidence des négations obligatoires (***ne pas payer de mine***). Les insertions fournissent aussi un aperçu de la variabilité syntaxique. Dans les exemples de la Table 4, le clivage (*c'est [...] que/ c'est [...] qui*) et la relativation sont repérées par l'insertion d'un relatif et classées manuellement.

A l'instar de ce que l'on observe avec le clivage, l'utilisation libre du nom de l'EPV en dehors de

| Catégorie de l'élément inséré | Exemple | ID occ. | ID % | LVC occ. | LVC % | IReflV occ. | IReflV % |
|---|---|---|---|---|---|---|---|
| ADJ | ***transfert** bancaire **effectué*** | 3 | 0.97 | 15 | 2.35 | 1 | 0.67 |
| ADV | ***fit** notamment **campagne*** | 136 | **44.16** | 27 | 4.23 | | |
| REL | ***transformations** que **subit*** | | | 3 | 0.47 | | |
| DETERMINANT : | | | | | | | |
| article | ***faire** un **tour*** | 88 | **28.57** | 424 | **66.46** | | |
| démonstratif | ***conclu** cette **union*** | | | 17 | 2.66 | | |
| indéfini | ***joue** aucun **rôle*** | 5 | 1.62 | 27 | 4.23 | | |
| numéral | ***effectué** deux **mandats*** | | | 12 | 1.88 | | |
| possessif | ***attirer** votre **attention*** | 23 | 7.47 | 49 | 7.68 | | |
| (SEMI-)AUXILIAIRE : | | | | | | | |
| être | ***en** est **arrivée*** | 6 | 1.95 | 19 | 2.98 | 119 | **79.33** |
| avoir | ***Il** aura **fallu*** | 13 | 4.22 | | | | |
| modal : *devoir, sembler,..* | ***il** peut **pleuvoir*** | 10 | 3.25 | | | | |
| NOM | ***traité** Euratom, **signé*** | | | 2 | 0.31 | | |
| PREP | ***mesures** à **prendre*** | 1 | 0.32 | 30 | 4.70 | | |
| PRONOM : | | | | | | | |
| démonstratif | ***prendre** cela **au sérieux*** | 2 | 0.64 | | | | |
| personnel | ***Il** me **semble*** | 5 | 1.62 | | | | |
| en | ***il y** en **a*** | 4 | 1.30 | | | 16 | **10.67** |
| y | ***s'y fier*** | | | | | 13 | 8.67 |
| le | ***il** le **faut*** | 4 | 1.30 | | | 1 | 0.67 |
| passif SE | ***match** se **dispute*** | | | 13 | 2.04 | | |
| t euphonique | ***Y a** -t -**il*** | 8 | 2.60 | | | | |
| TOTAL | | 308 | 100 | 638 | 100 | 150 | 100 |

TABLE 3 – Nature des insertions dans le cas d'une seule insertion. Les pourcentages les plus significatifs sont mis en gras.

celle-ci (Ex. 10) nous renseigne sur sa capacité à fonctionner sémantiquement de façon autonome et laisse supposer une EPV de type LVC.

**Ex. 10.** *Les deux pilotes déclassés **firent appel**, appel jugé en janvier 2007*

Les passivations potentielles sont détectées grâce au patron NOM-VB. La présence du verbe ÊTRE non annoté dans la fenêtre d'annotation indique une passivation avec auxiliaire. Le repérage des EPV passivées sans AUX se fonde sur la terminaison du verbe. La passivation en SE (*le **match** se **dispute***) est identifiée par l'insertion de ce pronom. Les occurrences trouvées ont fait l'objet d'une vérification manuelle.

Pour ces différents paramètres (passivation, relativation, clivage), on observe une moindre variabilité des ID que des LVC (Table 4). Les ID ont également davantage tendance à lexicaliser les déterminants (20 fois plus souvent) ou les négations (10 fois plus).

| type de variation | LVC occ. | LVC % | exemple LVC | ID occ. | ID % | exemple ID |
|---|---|---|---|---|---|---|
| passive AUX + VB | 59 | 4.3 | *son **rôle** est **joué*** | 5 | 0.3 | *Le **ton** a d'ailleurs été **donné*** |
| passive sans AUX | 135 | 9.9 | *le **rôle joué*** | 5 | 0.3 | ***compte-tenu*** |
| passive SE + VB | 45 | 3.3 | *La **finale** se **joue*** | 0 | 0 | |
| **TOTAL passives** | 239 | 17.6 | | 10 | 0.6 | |
| relative | 48 | 3.5 | ***tâche** qu'il a **accompli*** | 1 ? | 0.06 | ***question** que nous devons nous **poser*** |
| clivage | 2 | 0.1 | *c'est du bon **boulot** que vous m'avez **fait*** | 0 | 0 | |
| $\geq$ 1 det lexicalisé | 13 | 1 | ***avoir l'impression*** | 346 | 19.5 | ***tentent le tout pour le tout*** |
| négation lexicalisée | 1 ? | 0.1 | ***ne pas avoir de prix*** | 20 | 1.1 | ***ne pas tarir*** |
| TOTAL | 1357 | 100 | | 1777 | 100 | |

TABLE 4 – Certains aspects de variabilité et de figement des LVC vs ID sur l'ensemble du corpus

# 8 Analyse syntaxique de la variabilité

L'étude des insertions implique que les éléments en dehors de la fenêtre d'annotation ne sont pas pris en compte, comme l'adjectif dans l'exemple 11 :

**Ex. 11.** *les Byzantins [**remportent** une **victoire**] décisive*. (PARSEME)

Pour bénéficier d'une vue plus large des contraintes exercées sur les 29 EPV de type VB-NOM les plus fréquentes et n'ayant pas de déterminant figé, nous avons donc utilisé les relations de dépendance fournies dans le corpus, ce qui permet aussi une analyse plus aisée des insertions multiples (soulignées dans l'ex. 12)

**Ex. 12.** *La famille du cinéma français a [**rendu** vendredi à Paris un émouvant **hommage**] à l'un de ses plus brillants représentants* [...] (PARSEME)

Pour chaque expression testée, une seule occurrence de chaque type de dépendance du nom est comptabilisée. En effet, on s'intéresse à la capacité du nom à être modifié plutôt qu'au nombre de modifieurs [9]. On obtient 374 patrons de dépendances pour les noms inclus dans des ID et 434 dans des LVC, valeurs suffisamment proches pour permettre une comparaison. L'absence de dépendances est trois fois plus fréquente dans les ID que dans les LVC (48 % vs 13,6 %) et, dans une proportion similaire, le déterminant non lexicalisé est plus fréquent dans les LVC (30,4 % vs 7,5 %). Alors que le nom garde tout son sens dans les LVC, cela se produit rarement dans les ID (***faire office***) sans pour autant être impossible (***faire prisonnier***). Cela peut expliquer qu'il y ait plus fréquemment une absence de dépendances du nom dans les ID (***faire** *un grand **office***) que dans les LVC (***faire** une grande **révélation***).

En s'appuyant sur l'exploitation des dépendances, on peut :

- établir un classement des EPV de type VB-(DET)-NOM selon leur degré de variabilité
- identifier des comportements différents du NOM selon qu'il fait partie (ou non) d'une EPV,
- dégager les profils syntaxiques des variantes d'EPV de type VB-(DET)-NOM pour faciliter leur génération ou fusion automatique.

## 8.1 Classement des EP selon leur degré de variabilité

Tutin (2016) a étudié la variabilité des 30 EPV (ID et LVC) de type VB-(DET)-NOM les plus fréquentes en français. Les propriétés [10] les plus discriminantes sont la pluralisation du nom, les constructions relatives et passives (*le **rôle** est **joué** ; l'**attention** prêtée*). 2 EPV ont un comportement inattendu : ***avoir recours*** (LVC) s'avère moins variable que ***avoir** du **mal*** (ID) [11]. Mais Tutin (2016) souligne l'ambiguïté de cette EP [12] (***avoir mal / avoir** le **mal** de X*).

Nous avons cherché à comparer les niveaux de variabilité définis par Tutin (2016) pour certaines EP avec les tendances observables en corpus. Il s'agit d'une comparaison partielle faute de disposer des mêmes EP ou de ne pas en avoir un échantillon représentatif (moins de 10 occ.), d'où l'absence des

9. Par exemple, la phrase *Les femmes*[nsubj] *jouent au football*[nmod] *depuis la fin*[nmod] *du XIXe siècle en Angleterre*[nmod] aura pour dépendances associées au verbe JOUER : nsubj|nmod et non nsubj|nmod|nmod|nmod.

10. Chacune des 5 propriétés satisfaites accroît le niveau de variabilité d'une unité : pluralisation du nom possible, déterminant (ou son absence) non figé(e), relativation, passivation, nom modifiable par un adjectif.

11. Dans notre corpus : ***avoir** (du) **mal**, **avoir** (également) beaucoup de **mal**, n'**avoir** aucun **mal**, **avoir** plus **mal***.

12. Cette ambigüité est également présente dans ***poser question / poser** une **question*** quoique seule la seconde figure dans notre corpus. Pour ***jouer** un **rôle***, toutes les occurrences sont comptabilisées.

| Niveaux définis par Tutin (2016) → | Niveau 4 | Niveau 4 | Niveau 1 | Niveau 1 | Niveau 0 | Niveau 0 |
|---|---|---|---|---|---|---|
| EP | ***jouer rôle*** | ***poser question*** | ***faire partie*** | ***tenir compte*** | ***faire appel*** | ***donner lieu*** |
| Type | LVC | LVC | ID | ID | LVC | ID |
| VB-NOM | 79 % | 50% | 100 % | 71 % | 100 % | 100 % |
| NOM-VB | 21 % | 50% | | 29 % | | |
| 0 insertions | 9 % | 14 % | 93 % | 94 % | 94 % | 92 % |
| 1 insertion | 76 % | 36 % | 7% | 6 % | 6 % | 8% |
| 2 insertions | 9% | 36 % | | | | |
| > 2 insertions | 6% | 14 % | | | | |
| Nom variable | oui | oui | non | non | non | non |
| Passivation | oui | oui | non | oui | non | non |
| Relativation | oui | oui | non | non | non | non |
| nombre moyen par occurrence de types de dépendances différentes du nom (y compris absence de dépendance) | 0,39 | 0,57 | 0,03 | 0,29 | 0,09 | 0,17 |
| Total | 33 occ. | 14 occ. | 90 occ. | 17 occ. | 32 occ. | 12 occ. |

TABLE 5 – Comparaison de la variabilité observée en corpus avec les nivaux établis par Tutin (2016). Les cellules grisées indiquent des invariabilités. Le niveau 0 correspond à la variabilité minimale et le niveau 4 à la variabilité maximale.

niveaux 2 et 3 dans la Table 5. On y observe que les résultats sont cohérents pour les degrés extrêmes de (non-)variabilité mais moins marqués pour le niveau 1 puisque, dans notre corpus, l'expression ***faire partie*** a un comportement similaire à celui de ***faire appel***. Cette tendance est corroborée par le nombre de dépendances différentes associées au nom figurant dans l'EPV puisque ***faire partie*** montre une plus faible variabilité que les 2 expressions du niveau 0.

## 8.2 Dépendances de différents noms en EPV (ID) et hors EPV

Nous avons observé (section 7) que les dépendances du nom étaient différentes selon qu'il s'agissait d'un ID ou d'une LVC [13]. Lorsqu'un mot n'apparaît qu'au sein d'EPV (comme *mouron* dans ***se faire du mouron***), on le qualifie de *cranberry word* (Richter & Sailer, 2003). Ce phénomène étant rare, nous avons comparé les dépendances du nom lorsqu'il faisait partie du type d'EPV le plus contraint (ID) et lorsqu'il fonctionnait de façon autonome (hors EPV). Parmi les noms les plus fréquents dans les ID du corpus, *partie* et *objet* sont notamment présents dans les EPV ***faire partie/être de la partie*** et ***faire l'objet***. Lorsque ces noms font partie d'un ID, ils ont tendance (Table 6) à :

— moins souvent régir une préposition,
— moins souvent régir un modifieur adjectival,
— plus souvent régir un modifieur nominal.

S'il y a plus de modifieurs nominaux pour les noms figurant dans ces ID que lors d'un fonctionnement autonome, cela peut s'expliquer par leur structure syntaxique (que partagent 9 des 14 ID de la Table 2). En effet, ils requièrent un complément (pronom ou nom) : ***faire partie*** *de NOM,* ***faire l'objet*** *de NOM*.

Pour le nom *objet*, qu'il s'agisse ou pas d'une EP, la présence de déterminant se produit avec une fréquence assez similaire (quoique supérieure dans ce premier cas). Le déterminant est en effet lexicalisé dans l'EPV ***faire l'objet***. En revanche, l'association de *partie* avec un déterminant apparaît peu souvent dans une EPV compte-tenu de la sur-représentation de l'expression ***faire partie*** (94 %) interdisant sa présence.

13. La capacité du nom à être modifié ne fait pas partie des critères d'identification/classification des EP, dans le cas contraire l'intérêt de cette étude serait caduc.

| catégorie d'un élément régi par le nom | *partie* | | *objet* | | Exemples |
|---|---|---|---|---|---|
| | hors EPV | EPV | hors EPV | EPV | |
| det | 24.7 % | 2.4 % | 37 % | 43.6 % | *ils **seront** bien **de la partie** (EPV)*<br>*les deux parties de l' Empire* |
| nmod | 20.8 % | 82 % | 6.2 % | 35.2 % | *il **fait partie** du groupe (EPV)*<br>*la plus grande partie du personnel* |
| amod | 12.4 % | 4.8 % | 12.3 % | 1.4 % | *il **fait partie** des plus anciens (EPV)*<br>*en grande partie* |
| case | 16.5 % | 2.4 % | 17.2 % | 0 % | *ils **seront** bien **de la partie** (EPV)*<br>*en grande partie* |
| TOTAL | 232 occ. | 84 occ. | 37 occ. | 31 occ. | |

TABLE 6 – Dépendances des noms *partie* et *objet* en fonction de leur (non-)appartenance à une EPV. La terminologie *amod* (modifeur adjectival), *nmod* (modifieur nominal) et *case* (préposition) provient du tagset des *Universal Dependencies.*

## 8.3 Schémas de dépendance des différentes variantes syntaxiques

Nous cherchons ici à à établir des profils syntaxiques en dépendances pour différentes manifestations en surface des EPV. Cela permettra de générer ou normaliser automatiquement des variantes à partir de (ou vers) la forme canonique ou une autre variante. En utilisant un nouveau corpus parsé, nous pourrions ainsi faire des requêtes intégrant ces variantes pour accroître le nombre d'EPV détectées.

Nous nous sommes focalisés sur les 3 EPV (***jouer** un **rôle**, **signer** un **contrat /accord/partenariat**...), **jouer** la **finale***) offrant une variation syntaxique importante pour en décrire les régularités de construction. Parfois le verbe et le nom d'une EPV n'entretiennent pas, d'après les schémas de dépendances fournis, une relation de dépendance directe : dans ***signer** une quarantaine de **conventions***, le nom *quarantaine* s'interpose entre ***signer*** et ***conventions***. L'emploi de tels déterminants complexes s'observe 7 fois (***effectuera** une remarquable seconde partie de **saison**,...*).

L'identification de l'élément régisseur, des dépendances associées à celui-ci ainsi qu'à l'élément régi permettent d'identifier les différentes variantes syntaxiques. La Table 7 montre, pour chaque variante, la hiérarchie des éléments au sein de la chaîne de dépendance et leurs caractéristiques. Dans cette hiérarchie, on attribue la valeur 1 à l'élément régissant et cette valeur croît à chaque niveau de dépendances. A titre de comparaison, la fréquence des différentes variantes des 2090 occurrences de type VB-(DET)-NOM et NOM-VB dans le corpus s'avère relativement similaire.

| Type de variante | 3 EPV testées | | % corpus | élément | éléments régis | |
|---|---|---|---|---|---|---|
| | Exemples | Fréquence | VB-(DET)-N et N-VB | régissant | | |
| Niveau de dépendance | | | | 1 | 2 | 3 |
| forme canonique | *Il **signe** le **contrat*** | 85,9 % | 85,6% | **VB** | **NOM**-dobj | |
| passive avec AUX | *le **contrat** est **signé*** | 1,6 % | 2,8 % | **VB** | **NOM**-subjpass<br>+ auxpass-AUX | |
| passive sans AUX | *le **contrat signé*** | 6,6 % | 6,6 % | **NOM** | **VB**-acl | |
| passive en SE | *la **finale** se **joue*** | 2,9 % | 2,1 % | **VB**<br>+ se-dobj-PRON | **NOM**-nsubj | |
| infinitif | *a un **rôle** à **jouer*** | 1,2 % | 0,4 % | **NOM** | **VB**-acl | case (*à, de*) |
| clivage | | 0 % | 0,05 % | | | |
| relative | *l'**accord** qu'il a **signé*** | 1,6 % | 1,5 % | **NOM** | **VB**-acl :relcl | |

TABLE 7 – Différents schémas de dépendances : forme canonique, passivation, relative,... acl = *adjective clause* ; acl :relcl = *relative clause modifier* ; dobj = objet direct ; auxpass = auxiliaire passif.

# 9 Conclusion et perspectives

Ce travail confirme des comportements différents pour les EPV de type LVC et ID : les LVC ont une variabilité supérieure aux ID au niveau de la flexion nominale, de la relativation ou du clivage. La passivation y est aussi plus fréquente (17,6 % *versus* 0,6 %). De plus, les noms des LVC VB-NOM ont tendance à avoir une plus grande variété et quantité de dépendances que ceux des ID. Cependant, les noms employés dans les 2 ID examinés ont de 4 à 7 fois plus tendance à avoir des modifieurs nominaux qu'en dehors d'une EPV. Un examen à plus grande échelle sera nécessaire pour confirmer et affiner ce constat.
Les motifs de variabilité qui se dégagent pourront être mis à profit pour améliorer la détection d'EPV déjà connues. La typologie des insertions uniques et les schémas de variation syntaxique en dépendances permettront de générer ou normaliser automatiquement différentes variantes pour les EPV de type VB-(DET)-NOM. Comme il est peu probable que toutes les variations possibles d'une EPV donnée soient représentées dans notre corpus, il faudra recourir à d'autres ressources afin d'obtenir la fréquence d'apparition de ces variantes. Ces statistiques permettront d'attribuer aux variantes un caractère obligatoire, optionnel ou impossible : *il **attire** mon /*Ø/*deux **attention**(s) / mon **attention** est **attirée***. Pour les LVC, une étude conjointe des expressions contenant un même nom (ou un synonyme) en utilisant par exemple le Lexicoscope [14], serait sans doute judicieuse (*{**conclure / signer / sceller**} un {**accord / partenariat / alliance**, etc.}*).
Enfin, la recherche de dépendances interposées (***signer** une quarantaine de **contrats***) est un autre axe d'étude envisagé pour obtenir une vision globale de la variabilité dans l'objectif de favoriser la détection des EPV quelque soit leur forme de surface.

# Références


ABEILLÉ A. (1988). Light verb constructions and extraction out of NP in a tree ad-joining grammar. *Papers of the 24th Regional Meeting of the Chicago Linguistics Society*.

BALDWIN T., BENDER E. M., FLICKINGER D., KIM A. & OEPEN S. (2004). Road-testing the English Resource Grammar over the British National Corpus. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

CANDITO M., CONSTANT M., RAMISCH C., SAVARY A., PARMENTIER Y., PASQUER C. & ANTOINE J.-Y. (2017). Annotation d'expressions polylexicales verbales en français. In *TALN 2017*, Actes de TALN 2017, Orléans, France : Association pour le Traitement Automatique des Langues.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of TALN 2012*.

GROSS M. (1982). Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, **11**(2).

GROSS M. (1988). Les limites de la phrase figée. *Langages*, **23**(90), 7–22.

GROSS M. & SENELLART J. (1998). Nouvelles bases statistiques pour les mots du français. In *Journée d'Analyse Statistique des Données Textuelles (JAD T)*, p. 335–349, Nice.

JACKENDOFF R. (1997). The architecture of the language faculty. *Linguistic Inquiry Monographs*.

JACQUEMIN C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press.


14. http://phraseotext.u-grenoble3.fr/lexicoscope/

LAPORTE E. (1988). Reconnaissance des expressions figées lors de l'analyse automatique. *Langages*, **23**(90), 117–126.

LICHTE T., PETITJEAN S., SAVARY A. & WASZCZUK J. (2017). *Lexical encoding formats for multi-word expressions : The challenge of "irregular" regularities"*, In Y. PARMENTIER & J. WASZCZUK, Eds., *Representation and Parsing of Multiword Expressions*, p. 79–111. Language Science Press, à paraître.

NISSIM M. & ZANINELLO A. (2013). Modelling the internal variability of multiword expressions through a pattern-based method. In *ACM Transactions on Speech and Language Processing , Special issue on Multiword Expressions*, volume 10.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* : European Language Resources Association (ELRA).

NUNBERG G., SAG I. A. & WASOW T. (1994). Idioms. *Language*, **70**, 491–538.

PRZEPIÓRKOWSKI A., HAJNICZ E., PATEJUK A. & WOLIŃSKI M. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, p. 83–91, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

RICHTER F. & SAILER M. (2003). Cranberry words in formal grammar. In C. BEYSSADE, O. BONAMI, P. CABREDO HOFHERR & F. CORBLIN, Eds., *Empirical Issues in Formal Syntax and Semantics 4*, volume 4, p. 155 – 171. Presses de l'Université de Paris-Sorbonne, Paris.

SAG I., BALDWIN T., BOND F., COPESTAK A. & FLICKINGER D. (2002). Multiword expressions : A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, p. 1–15, Mexico City,Mexico.

SAVARY A. (2008). Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, p. 1–53.

SAVARY A. & JACQUEMIN C. (2003). Reducing Information Variation in Text. *LNCS*, **2705**, 145–181.

SAVARY A., RAMISCH C., CORDEIRO S., SANGATI F., VINCZE V., QASEMIZADEH B., CANDITO M., CAP F., GIOULI V., STOYANOVA I. & DOUCET A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, p. 31–47, Valencia, Spain : Association for Computational Linguistics.

SHEINFUX L. H., GRESHLER T. A., MELNIK N. & WINTNER S. (2017). *Verbal MWEs : Idiomaticity and flexibility*, In Y. PARMENTIER & J. WASZCZUK, Eds., *Representation and Parsing of Multiword Expressions*, p. 5–38. Language Science Press, à paraître.

TUTIN A. (2016). Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French : a corpus based study. In *PARSEME COST Action. Relieving the pain in the neck in natural language processing : 7th final general meeting*, Dubrovnik, Croatia.

VILLAVICENCIO A., KORDONI V., ZHANG Y., IDIART M. & RAMISCH C. (2007). Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 1034–1043, Prague, Czech Republic.