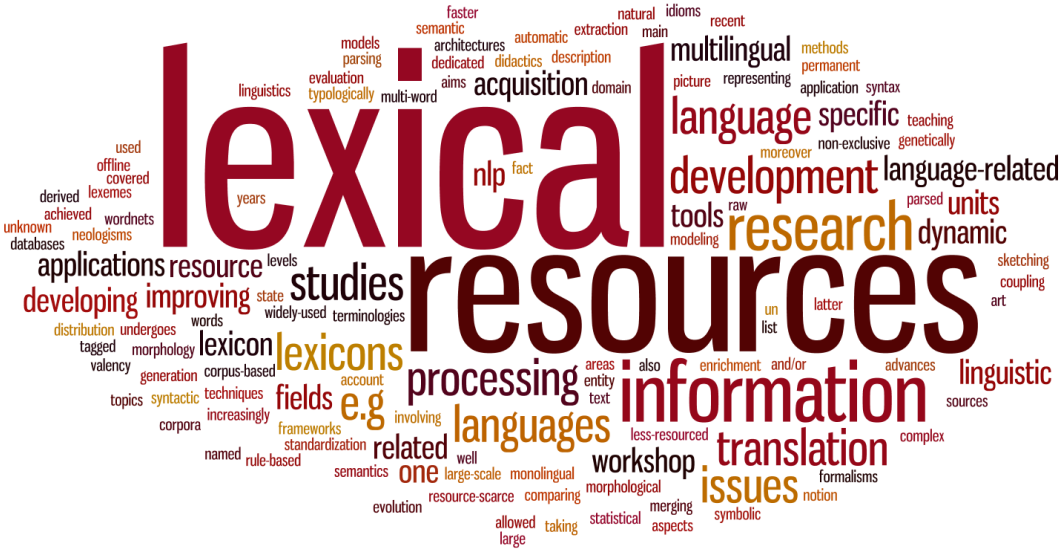
First International Workshop on Lexical Resources

Benoît Sagot

First International Workshop on Lexical Resources

An ESSLLI 2011 Workshop



Ljubljana, Slovenia — August 1–5, 2011

First International Workshop on Lexical Resources

Program committee

Workshop chair

Benoît Sagot - Alpage, INRIA & Université Paris 7, France

Members

Afra Alishahi - Saarland University, Germany
Marianna Apidianaki - INRIA, France
Núria Bel - Universitat Pompeu Fabra, Spain
Francis Bond - Nanyang Technological University, Singapore
Paul Buitelaar - National University of Ireland - Galway, Ireland
Nicoletta Calzolari - Istituto di Linguistica Computazionale, Italy
Béatrice Daille - Université de Nantes, France
Laurence Danlos - Université Paris 7, France
Helge Dyvik - Bergen University, Norway
Tomaž Erjavec - Jožef Stefan Institute, Slovenia
Christiane Fellbaum - Princeton University, USA
Jennifer Foster - Dublin City University, Ireland
Yoav Goldberg - Ben Gurion University of the Negev, Israel
Shu-Kai Hsieh - National Taiwan University, Taiwan
Philippe Langlais - Université de Montréal, Canada
Éric Laporte - Université Paris-Est Marne-la-Vallée, France
Linlin Li - Saarland University, Germany
Piet Mertens - Katholieke Universiteit Leuven, Belgium
Karel Pala - Masaryk University, Czech Republic
Stelios Piperidis - Institute for Language and Speech Processing, Greece
Adam Przepiórkowski - Polish Academy of Sciences, Poland
Francis Tyers - Universitat d'Alacant, Spain
Duško Vitas - University of Belgrade, Serbia
Piek Vossen - Vrije Universiteit Amsterdam, Netherlands
Pierre Zweigenbaum - LIMSI, France

First International Workshop on Lexical Resources

Table of contents

Introduction

Lexical resources are one of the main sources of linguistic information for research and applications in Natural Language Processing and related fields. In recent years advances have been achieved in both symbolic aspects of lexical resource development (lexical formalisms, rule-based tools) and statistical techniques for the acquisition and enrichment of lexical resources, both monolingual and multilingual. The latter have allowed for faster development of large-scale morphological, syntactic and/or semantic resources, for widely-used as well as resource-scarce languages. Moreover, the notion of dynamic lexicon is used increasingly for taking into account the fact that the lexicon undergoes a permanent evolution.

WoLeR 2011, the First International Workshop on Lexical Resources, aimed at sketching a large picture of the state of the art in the domain of lexical resource modeling and development. It was also dedicated to research on the application of lexical resources for improving corpus-based studies and language processing tools, both in NLP and in other related fields, such as linguistics, translation studies, and didactics.

WoLeR 2011 was an ESSLLI 2011 workshop. The European Summer School in Logic, Language and Information (ESSLLI) has been organized every year by the Association for Logic, Language and Information (FoLLI) in different sites around Europe. The main focus of ESSLLI is on the interface between linguistics, logic and computation. In 2011, ESSLLI was held in Ljubljana, Slovenia and was organized by the Slovenian Language Technologies Society (SDJT), the Jožef Stefan Institute (IJS) and The Faculty of Mathematics and Physics (FMF) in Ljubljana, Slovenia. Chair of the Program Committee was Makoto Kanazawa (National Institute of Informatics, Japan), and Chair of the Organizing Committee was Darja Fišer (U. of Ljubljana).

The invited talk at WoLeR 2011 was Caroline Sporleder, head of the "Computational Modelling of Discourse and Semantics" group at the Computational Linguistics and Phonetics Department, Saarland University located in Saarbücken, Germany. The title of her talk was *Towards Large-Scale Lexical Semantic Resources*. In the name of all participants, I would like to thank her very warmly for her rich and inspiring talk.

All 22 submissions were reviewed by at least two members of the Program Committee, and 12 were accepted for an oral presentation. They cover a wide range of topics from the point of view of lexical resources, from morphology to syntax and semantics, but also didactics, and a large range of languages, such as Arabic, Dutch, English, French, Spanish, as well as contemporary and 19th century Slovene.

I would like to thank all members of the Program Committee who did an fantastic job in reviewing the submitted papers and providing the authors with invaluable comments. I would like to thank also Alpage, a joint team at INRIA and Université Paris Diderot – Paris 7 (France) and the research project EDyLex (ANR-09-CORD-008), funded by the Agence Nationale de la Recherche, for their financial support, as well as FLaReNet for its endorsment. Finally, special thanks go to the ESSLLI 2011 organizers, and in particular Darja Fišer, who have dealt with all the logistics and helped us focusing on the topic of WoLeR: lexical resources.

Benoît Sagot
Paris, August 2011

Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus

A. Balvet*, L. Barque, M.-H. Condette*, P. Haas*, R. Huyghe***, R. Marín*, A. Merlo***

*STL, CNRS UMR 8163, **LDI, CNRS UMR 7187, ***EILA
U. Lille 3, U. Paris 13, U. Paris 7
prenom.nom@univ-lille3.fr, pnom@ldi.univ-paris13.fr, pnom@eila.univ-paris-diderot.fr

Abstract

ANR-funded Nomage project aims at describing the aspectual properties of deverbal nouns taken from a corpus, in an empirical way. It is centered on the development of two resources: a semantically and syntactically annotated corpus of deverbal nouns based on the French Treebank, and an electronic lexicon, providing descriptions of morphological, syntactic and semantic properties of the deverbal nouns found in our corpus. Both resources are presented in this paper, with a focus on the comparison between corpus data and lexicon data.

1. Introduction

From a theoretical standpoint, the works of (Lees, 1960), through (Chomsky, 1970) and (Grimshaw, 1990), provide a laying ground for our description of deverbal nouns' properties, though these works focus mainly on morphological and syntactic aspects. We elaborate on this theoretical framework, by providing fine-grained descriptions of the morphological, syntactic, semantic (more precisely aspectual) properties of deverbal nouns in an empirical way. In this paper, after a brief revision of related work (section 2.), we present the Nomage corpus and the semantic annotation process applied to deverbal nouns (section 3.). We then present the structure and content of our lexicon, which describes the deverbal nouns extracted from our corpus, alongside the morphologically-related verbs we manually associated to each of these nouns (section 4.). Since, in our project, the description of deverbal nouns is carried out by means of two different methods, in the last section we confront annotations taken from the corpus with those taken from the lexicon (section 5.).

2. Related work

Leaving aside Verbaction (Tanguy and Hathout, 2002), an xml database of nominalizations paired with their verbal bases, the resource we present here is, as far as we know, the first attempt to semantically annotate both a corpus and a lexicon of French deverbal nouns.

Similar resources exist for other languages, particularly for English and Spanish. For English, the most relevant resource is NOMLEX, a lexicon of English deverbal nominalizations containing 1,025 entries (Macleod et al., 1998). It is mainly focused on argument structure: the allowed complements of nominalizations are described and linked to their corresponding verbal arguments. NOMLEX-PLUS (Meyers et al., 2004), an integral part of the NomBank project (Meyers, 2007), is an extension of NOMLEX. It includes 7,050 additional entries: 4,900 for verbs' nominalizations, 550 for adjectives' nominalizations, and 1,600 corresponding to other argument-taking nouns.

For Spanish, one can cite AnCora-Nom (Peris et al., 2010), a lexicon of 1,655 lexical entries corresponding to the different deverbal nominalizations appearing in the annotated corpus Ancora-Es (Taulé et al., 2008). Ancora-Nom not only includes information on argument structure, like NOMLEX, but also on lexical aspect.

3. The corpus

In this section, we outline the main features of the electronic corpus we use, the French Treebank, and we describe the deverbal noun candidates' extraction process. Then, we proceed by describing our semantic annotation protocol.

The French Treebank is a 1 million words corpus of newspaper articles taken from *Le Monde*. It provides several levels of linguistic annotations: simple and compound tokenization, lemmas, part-of-speech tags augmented with morphological information, together with constituent boundaries and syntactic functions for half of the corpus (Abeillé et al., 2003).

Based on morphological cues (suffixes: -ion, -age, -ment, etc.), we extracted over 10,000 nominalization candidates (simple tokens only) from the functionnaly-annotated half of the French Treebank. After close inspection, only 4,042 candidates were considered in the course of the project: all nouns that were not syntactic heads (e.g. *un permis de* ***construction*** *(a construction permit)* versus *la* ***construction*** *européenne (the European integration)*) of a NP were discarded, because of their incompatibility with the transformation tests we used for the semantic annotation process (see below, section 3.2.). Moreover, some nominalizations stem from an adjectival base, and not a verbal one: e.g. INDULGENCE stems from INDULGENT, but our project aims exclusively at deverbal nouns. The Nomage project is dedicated precisely to the study of the inheritance of semantic and aspectual features from the verbal bases, thus some of the extracted candidates, which were possible converted nouns, were also discarded. Even though a link to a verbal lexeme can be found, the directionality of the inheritance relationship cannot be clearly established; this includes cases such as VOYAGE (*travel*, noun) and VOYAGER (*travel*, verb). Finally, some amount of noise is attributable to the extraction process itself. Morphological cues in themselves do not discriminate between true nominaliza-

tions and false-positives: items such as SARCOPHAGE (*sarcophagus*) had to be filtered-out, based either on an automatic filtering ("stop-list" lookup) or a manual process.

3.1. Aspectual annotation of deverbal nouns

One of the central methodological features of project Nomage is that the semantic descriptions rely on the application of transformation tests, carried out by "naive" annotators and not on forged examples. These tests were devised so as to highlight a selection of semantic properties for each candidate: its aspectual structure, together with its mass/count status. We wish to emphasize here that the transformational tests were intentionally devised so as to be applied by native speakers that had received no training in linguistics. These annotators were not aware of the fine-grained semantic and aspectual distinctions we were trying to describe, but rather they were simply asked to assess whether each transformation yielded acceptable sentences or not.

Originally, we had planned to implement a cross-annotation process for each candidate, in order to provide minimal inter-annotator agreement (kappa) assessment. Unfortunately, due to a lack of available annotators, this methodology had to be abandoned: up to 7 "naive" annotators were hired for this project, some of them at different points in time, working on partially intersecting annotation batches, while a minimum of 10 *distinct* annotators would have been required. Moreover, due to data integrity issues, part of the candidates had to be manually corrected by researchers (and thus far from "naive") associated to the project. Therefore, it is not possible to provide inter-annotator agreement scores for our data.

3.2. Using transformational tests to assess semantic properties

The transformational tests were voluntarily presented in an unstructured manner to the annotators, so as to avoid any implicit theory-forming on their part. We present below the semantic annotation of nominalization ÉVALUATION, based on our methodology.

*L'**évaluation** faite selon les critères du BIT (Bureau International du travail) n'est pas plus rassurante. [The evaluation carried out by the BIT is not more reassuring.]*
T1_Plusieurs : yes : → *Plusieurs évaluations* **T2_Avoir lieu :** yes → *L'évaluation qui a eu lieu hier* **T3_Éprouver/ressentir :** no **T4_Un peu de :** no **T5_Durer x temps :** yes → *L'évaluation qui a duré 2 jours* **T6_Se trouver (qq part) :** yes → *L'évaluation qui se trouve sur ton bureau* **T7_Effectuer/procéder :** yes → *L'évaluation effectuée hier* **T8_État de :** no **T9_Se dérouler :** yes → *L'évaluation qui s'est déroulée hier* **T10_Card :** yes → *Deux évaluations*

Table 1: Semantic annotation of *évaluation*

Our tests allow us to uncover two main semantic features: mass/count status, and aspectual structure. Annotators had to assess whether the original determiner could be replaced by *plusieurs* 'several' (test 1), *un peu de* 'some' (test 4) or by a cardinal determiner (test 10). Here, tests 1 and 10 yield a positive outcome, while test 4 is impossible, which allows us to categorize ÉVALUATION as a count noun.

As for aspectual properties, *avoir lieu* 'happen/hold' (test 2) and *effectuer/procéder* 'complete/perform' (test 7) are meant to identify whether the candidate has an **event** reading. Here, it is precisely the case: both tests can be applied. In addition, "se dérouler" (test 9) indicates that the considered noun is a durative event. Other tests are aimed at non-event readings: tests such as *éprouver/ressentir* 'feel' (test 3) and *état de* 'to be in a state of' (test 8)[1] allow us to identify **state** readings. Here, this occurrence of ÉVALUATION is not compatible with these latter tests, which is, in itself, a confirmation of its event reading. Finally, *se trouver (qq part)* 'to stand/be located at' is meant for capturing **object** readings.

3.3. Test outcomes and semantic categorization

Test outcomes on our 4,042 items are interpreted so as to yield 3 classes: EVT (events), ETAT (states) and OBJET (objects). In order to be categorized as a state, a candidate must exhibit at least one positive outcome for tests 3 or 8. For objects, only test 6 is considered, while for events a candidate must yield one positive outcome for test 2. Therefore, even though our tests may appear partly redundant, this is intentional, as some tests are considered as more generic and others more specific. In the case of events for instance, test 2 is more generic than test 9, it is thus more discriminating: "avoir lieu" allows us to distinguish event and non-event readings, while test 9 allows us to further specify an event subclass. Moreover, this design serves as a rough control mechanism so as to avoid inconsistencies in annotations: for example a positive outcome to test 9 is supposed to entail a positive outcome for test 2. Annotations that do not follow this pattern are easy to spot and are put under close scrutiny in the final validation process. As for test 5, it is used along with test 9 to discriminate a certain subclass of events –the durative ones as opposed to the punctual ones. But test 5 is also valid for states[2] and in some cases may help categorize them. As can be seen in table 2, the conjunction of different test outcomes is used to yield "inferred" semantic classes, which will be compared to hand-coded semantic classes in the lexicon, in section 5. Examples (1a) through (1c) and table 2 give an illustration of the semantic classes that can be associated to each occurrence, based on their respective test outcomes, as coded by our naive annotators.

(1) a. *L'**évaluation** faite selon les critères du BIT (Bureau International du travail) n'est pas plus ras-*

[1] For this test, the sequence "état de" has to be inserted between the candidate and its determiner : * *L'**état d**'évaluation faite selon le BIT* ...

[2] Test 5, to some extent, is also valid for objects (e.g. *Sa télé a duré 2 mois avant de tomber en panne*) but has not the same interpretation.

surante. [The evaluation carried out by the BIT is not more reassuring.]

b. *Il s'agit de produits récupérés à l'état liquide dans les **installations** de traitements des gaz. [This refers to liquid-state products recovered from gas processing facilities.]*

c. *Le **mécontentement** est de plus en plus grand en Pologne à la suite des fortes hausses des prix du gaz, de l'electricité et de l'eau chaude appliquées au début de l'année. [Discontent grows in Poland following a sharp increase in gas, electricity and hot water prices.]*

	(1a)	(1b)	(1c)
2. *Avoir lieu*	**yes**	no	no
3. *Éprouver*	no	no	**yes**
5. *Durer x temps*	**yes**	no	**yes**
6. *Se trouver*	**yes**	**yes**	no
7. *Effectuer/procéder*	**yes**	no	no
8. *État de*	no	no	no
9. *Se dérouler*	**yes**	no	no
Inferred class	EVT or OBJET	OBJECT	STATE

Table 2: Interpretation of aspectual test outcomes

As can be seen, based on test outcomes (table 2), the occurrence of EVALUATION in (1a) has two related meanings, an action and its result, that can be co-predicated in the same sentence (Pustejovsky, 1995; Godard and Jayez, 1996; Milicévic and Polguère, 2010).

4. The lexicon

4.1. A lexicon entry

The Nomage lexicon describes each deverbal noun from our corpus (amounting to 746 nominal lexemes)[3], as well as their verbal base (679 verbal lexemes). Each nominal lexeme is associated with an aspectual class and a semantic argument structure. Note that the aspectual class is not attributed to lexemes according to the results of the tests applied to their occurrences in the corpus (see section 3. above) but following a classical method that will be explained in section 4.1.2. below. We emphasize here that our goal is precisely to contrast two aspectual annotation methodologies.

Tables 3, 4 and 5 below illustrate the kind of information that can be found in the Nomage lexicon, with the description of AMÉNAGEMENT#1 and its verbal counterpart AMÉNAGER#1. As illustrated in table 3, AMÉNAGEMENT#1 has two arguments (X and Y) and denotes an accomplishment (i.e. a durative event).[4]

[3]These 746 nominal lexemes correspond to the 4,042 tokens in the corpus. The average number of examples per lexemes is thus 5.5.

[4]The aspectual classes assigned to each lexeme are based on a finer-grained ontology than the habitual three classes (EVT, STATE, OBJECT). In our lexicon, we distinguish for instance durative events from non durative ones, and telic from non telic ones (see below section 4.1.2.).

id	45
Lexeme	AMÉNAGEMENT#1
Argument structure	$\sim$ de Y par X
Aspectual class	ACC
Occurrences in the FT	{id:1794 ; id:1929}
Verbal base	id:44

Table 3: Description of noun AMÉNAGEMENT#1

Alongside the information given above, each entry points to a verbal source. It is thus possible to have access to a description of the verbal lexeme AMÉNAGER#1 through the nominal one AMÉNAGEMENT#1. As can be seen in table 4, the verb's argument structure and aspectual class are also described in our lexicon.

id	44
Lexeme	AMÉNAGER#1
Argument structure	X $\sim$ Y
Aspectual class	ACC

Table 4: Description of verb AMÉNAGER#1

Finally, each entry is associated with its corresponding occurrences in the original corpus, and the actual realization of the lexeme's arguments (table 5).

id	1794
Deverbal	id:45
Occ.	Tout ce travail préparatoire sera fondamental pour l'**aménagement** universitaire au cours des cinq prochaines années.
Réal. Arg.	X:$\emptyset$, Y:adj. rel.

Table 5: An occurrence of AMÉNAGEMENT#1

4.1.1. Argument structure

In our lexicon, we describe the semantic arguments of each nominal and verbal predicates in a systematic manner. By semantic arguments we mean the required participants in order to define the state of affairs denoted by the considered predicate (Mel'čuk, 2004a). Semantic arguments are represented by variables (X, Y, Z), as can be seen in the description of AMÉNAGEMENT#1, which is associated with two arguments X and Y. The Dicovalence lexicon (Van den Eynde and Mertens, 2003) frequently helped us to identify the semantic arguments of verbal predicates, which are generally also those of the corresponding nominal predicate. This is the case for AMÉNAGER/AMÉNAGEMENT: X represents in both cases the "agent" and Y the "undergoer". Each lexeme is associated with a description of the surface realization of its semantic arguments in the corpus[5] (Mel'čuk, 2004b). Lexeme AMÉNAGEMENT#1 occurs for example in the following sentences of the corpus :

[5]Note that not all possible realizations of a given semantic argument structure are described: we only consider the realizations found in our corpus.

(2) a. *Tout ce travail préparatoire sera fondamental pour l'**aménagement** universitaire au cours des cinq prochaines années.* X:∅, Y:adj. rel. (cf. table 5 above)

b. IBM devient ainsi actionnaire de Dassault systèmes à hauteur de 10% et assure la **commercialisation** de ses logiciels Catia. X:∅, Y:adj. rel., Verbe Support= X assurer det N

In sentence (2a), argument X of AMÉNAGEMENT#1 is not realized, while argument Y is realized by a relational adjective (*universitaire*). Note that arguments that are syntactically dependent from the light verb of a nominalization are also described: for example, semantic argument X of COMMERCIALISATION is realized as the subject of the light verb *assurer* in sentence 2a.

4.1.2. Aspectual class

We follow a classical approach to the description of the aspectual class of the deverbal nouns in our lexicon. We use aspectual tests, taken from the literature, in order to characterize their semantic and aspectual properties. In contrast, as has been shown above, the attribution of an aspectual class to each occurrence taken from our corpus was based on a set of transformational tests meant to be applied by "naive" annotators in the original context. We give a comparison between these annotation methods in section 5.

The first four labels retained are taken from Vendler's classification of verbs (1967), with slight adaptations, particularly by using the feature [+/- culminating], and extended to the nominal field. Lexemes of the states class (ETAT) denote non dynamic situations (e.g. POSSÉDER, ADMIRATION, etc.). On another branch of the aspectual ontology, lexemes of the activities class (ACT), such as MANIFESTER and PROMENADE denote dynamic, durative and non culminating situations. Accomplishments (ACC), such as RÉPARER and DÉMÉNAGEMENT, describe dynamic, durative and culminating situations. Finally, lexemes of achievement type (ACH) denote dynamic and culminating but non durative situations (e.g. ADOPTER and ACQUISITION).

The aspectual descriptions in our lexicon rely on original classes, as we have frequently observed that some lexemes do not match any of the simple classes mentioned above, but seem rather to constitute intermediate categories: thus, between achievements and states, we have proposed "stative achievements" (ACH-ETAT) which react positively to some tests dedicated to achievements but also to some tests accepted for states – particularly tests of duration, when these tests concern a resultant state. This class is dedicated to items such as EMPRISONNEMENT which denote an achievement (the sending to prison) followed by a state that lasts until the end of the process (the coming out of prison). In the same way, we propose "stative accomplishments" (ACC-ETAT) which describe an accomplishment followed by a state. This class encompasses cases such as INVASION which refers to the durative action of the invasion of a territory and to the state of occupation of the invaded territory. We have also introduced "accomplishments-activities" (ACC-ACT), which constitute an intermediate class between the ACT and the ACC, and denote activities of which each step could be considered as the final stage. This class comprises items such as REFROIDIR, RÉTRÉCISSEMENT, etc. This category is also known under the noun of "degree-achievement" (Dowty, 1979). The classes we have just presented apply at the same time to verbal and nominal lexemes. However, the existence of semantic idiosyncrasies in the nominal field has made us consider several new aspectual categories so as to label our nominalizations more finely.

More precisely, for the class of activities, we've had to add a label in the nominal field so as to take into account the fact that verbs of activity (e.g. JARDINER, SE PROMENER, MANIFESTER) do not yield a homogeneous class of nominalizations (Flaux and Van de Velde, 2000; Haas et al., 2008; Heyd and Knittel, 2009). Indeed, the opposition massive / countable distinguishes, at the aspectual level, two types of nouns derived from verbs of activity: countable nominalizations (e.g. PROMENADE) and massive nominalizations (e. g. JARDINAGE). From the aspectual point of view, all these nouns describe dynamic, durative and non culminating situations, but only count nouns denote actions which are temporally delimited, i.e. events (Haas and Huyghe, 2010). We keep the ACT label for these deverbal activity count nouns, which are statistically the most representative of the category, whereas their massive counterparts are labeled HAB (for "habitude"), because they can denote routine activities (Barque et al., 2009).

Another particularity of nouns is that, contrary to verbs, they can denote objects, and in this case they are devoid of any aspectual features. This property is known for the nominalizations that express the result of an action (Grimshaw, 1990), but it can be extended to a wider set of nominalizations. So we consider the existence of a class called OBJET, in which we group together nouns that denote material objects (e. g. CONSTRUCTION), nouns that denote objects with an informational content (e.g. AFFIRMATION), and nouns that denote entities which induce a psychological state (e.g. OBSESSION). Finally, we have used complex classes that include nominal lexemes which are likely to denote a situation and/or an object (Pustejovsky, 1995; Godard and Jayez, 1996; Milicévic and Polguère, 2010). These lexemes can receive co-predication, as in *Son exposé fut long et ennuyeux*, where *long*, which qualifies the presentation course and progress, applies to the "accomplishment" aspect of EXPOSÉ, whereas *ennuyeux*, which qualifies the informational content of the presentation, applies to the OBJET meaning. Such a case receives the ACC•OBJET label.

The tests for assigning an aspectual class to verbs are well known in the literature. But the aspectual properties of nouns have been less studied, so we've had to adapt the classical verb-oriented tests to this class of lexical units. The set of these tests, which are presented in detail in the documentation of the lexicon (written in French), is available at the following address: `http://nomage.recherche.univ-lille3.fr/` (attached in the "délivrables" part of the site).

Table 6 summarizes the different aspectual classes at-

tributed to each entry (nominal or verbal) in the lexicon.

Verbal classes	ACC, ACC-ETAT, ACH, ACH-ETAT, ACT, ACT-ACC, ETAT
Nominal classes	verbal classes + ACC•OBJET, ACH•OBJET, HAB, OBJET

Table 6: Aspectual classes in the lexicon

Table 11 shows the correspondance between the fine-grained classification used in the lexicon and the more general classification used in the corpus.

Corpus aspectual classes	lexicon aspectual classes
EVT	ACC, ACT, ACT-ACC, ACH, (ACC/ACH)-ETAT, (ACC/ACH)•OBJET
ETAT	ETAT, ACC-ETAT, ACH-ETAT
OBJET	OBJET, ACH•OBJET, ACC•OBJ

Table 7: Two sets of aspectual classes

4.2. Polysemy ratio

Table 8 shows the overall polysemy ratio of nominal and verbal forms from our lexicon.

Nominal lexemes	746
Nominal forms	656
Nominal polysemy ratio	1.14
Verbal lexemes	679
Verbal forms	648
Verbal polysemy ratio	1.04

Table 8: Polysemy ratio in the Nomage lexicon

The low polysemy ratio (1.14 lexemes by entry) can be explained by the fact that our corpus is relatively small (500,000 words) and specialized (newspaper articles). For deverbal nouns, polysemy comes from two main sources: it can either be inherited from the verbal base, or it can be attributed to the noun itself. For example, in the sentences below, each PROMOTION lexeme derives from two distinct PROMOUVOIR verbal lexemes.

(3) a. *C'est arrivé après sa **promotion** au poste de directeur financier.* (la personne X PROMOUVOIR#1 l'individu Y au poste Z → PROMOTION#1 de Y à Z accordée par X)

b. *Chirac va faire la **promotion** de son livre en plein marasme judiciaire.* (la personne X PROMOUVOIR#2 Y → PROMOTION#2 de Y par X)

In our lexicon, the other source of polysemy is mostly attributable to metonymy links that can be observed between an action and one of its participants or between an action and its result (Bisetto and Melloni, 2008). For instance, in our lexicon we describe two lexemes INSTALLATION, one denoting the fact of installing something, the other the result of the process (the installed thing).

5. Analysing data

5.1. Suffixes and aspectual classes in the lexicon

From a morphological point of view, one of the main descriptive and theoretical issues is the relationship between the aspectual class of a nominal lexeme and its suffix. The table below is a census of the different semantic class[6] → suffix mappings in our lexicon.

	EVT	STATE	OBJ	HAB	total
-ade	6	-	-	-	6
-age	45	2	7	2	56
-ance/-ence	10	19	8	2	39
-ée	13	2	3	-	18
-ion	336	36	61	12	445
-ment	133	12	14	3	162
-ure	11	1	6	2	20
total	554	72	99	21	746

Table 9: Distribution of aspectual classes by suffix

As can be seen, the most productive suffix is -ion (60.5%), followed by -ment (20.7%) and -age (7.6%). These results conform to those given by Tanguy & Hathout (2002). Regarding aspectual classes, events are the most frequent (75.3%) class, followed by objects (13.5%) and states (9.8%).
As for the relationship between suffix and aspectual class, we can notice that:

- -ance/-ence is the only suffix with less than 50% of events cases; this suffix also has the strongest tendency to combine with states. This result amends the rather widespread idea (Gaeta, 2002) that suffix -ance/-ence is only compatible with states. Our results show that, though it is true this suffix has a marked preference for states, it also combines with other aspectual classes (Dal and Namer, 2010).
- -age, -ée, -ment and -ion suffixes behave in similar fashions: between 70% and 80% of words bearing these suffixes are events.
- -ure offers fewer cases of events (55%); it is also the suffix which has the strongest tendency to combine with objects (30%).

Nevertheless, if we compute 95% confidence intervals with an error-rate of 0.05 based on figures of absolute frequencies over 100 occurrences, the size of the intervals is seldom under 7%. For instance, if we filter-out low-frequency suffix-aspectual class distributions and keep only those suffixes over 100 occurrences, the confidence interval for -ion as an EVT (336 occurrences) is [71.5;79.5]. For -ment, as an EVT (133 occurrences), it is [76.2;88][7]. Therefore, the

[6] As illustrated in table 11, we use two distinct set of aspectual classes : a fine-grained one to classify the lexemes and a more general one to classify occurrences of deverbal nouns in the corpus. The class HAB is the only one that can be generalized as EVENT, STATE or OBJECT.

[7] The 95% confidence intervals were computed based on a standard margin of error, following the function: Po ± 1.96 x

size of the confidence intervals is an indication that these figures are to be taken with extreme caution and should be computed on larger sets of data for higher confidence thresholds. Interestingly, χ^2 scores computed on these data show that the only suffix for which the null hypothesis should be discarded is -ance/-ence as a STATE[8]. Therefore, a strong connection between this suffix and the STATE class cannot be attributed to mere chance.

5.2. Verb → Noun inheritance of semantic properties

The main issue in this project is to assess whether a deverbal noun inherits (part of) the semantic and aspectual properties from the associated verbal form or not. In order to address this, we have assigned an aspectual class to each verb and noun described in our lexicon (see 4.1.2.), which enables us to compare and analyze matches and discrepancies between verbal and nominal domains. Our data indicate a perfect match between verbal and nominal aspectual class in around 67% of cases (492 perfect matches out of 737 verb-noun pairs). The remaining 245 verb-noun pairs exhibit at least some degree of discrepancy. Two main cases appear:

1. verbs and their nominalizations belong to two different classes entirely;
2. verbs and their nominalizations belong to slightly different classes.

5.2.1. Total verb-noun aspectual discrepancy

This case represents 73% (178 cases out of 245) of all mismatches, of which at least a partial explanation can be found in the existence of OBJECT classes for nouns, which by definition have no counterpart in the verbal domain. In this case, nominalizations do not denote an abstract situation (ACT, ACC, ETAT, etc.) but rather an object devoid of all aspectual properties. Around 55% of total discrepancies fall in this category (98 out of 178), for example: AGGLOMÉRER (ACC) → AGGLOMÉRATION (OBJET). The same holds for the HAB (routine activities) class for the nominal domain, which represents around 9% of the total discrepancy cases, e.g. RÉSISTER (ACT) → RÉSISTANCE (HAB). The remaining 64 verb-noun pairs (over 35%) are cases where the observed verb-noun aspectual mismatch cannot be explained by the existence of a class restricted to nouns: in some cases, only a slight discrepancy can be observed, e.g. INTERVENIR (ACC) → INTERVENTION (ACT) (in both cases we are dealing with durative events). In other cases, a major discrepancy can be observed, between the verbal and nominal domains, e.g. SOUFFRIR (ACT) → SOUFFRANCE (ETAT) (shift from dynamic to stative situation).

5.2.2. Partial verb-noun aspectual discrepancy

67 verb-noun pairs out of our 178 aspectual discrepancy cases are only partial mismatches. One of the causes for such mismatches is simply the overall discrepancy between verbal and nominal aspectual ontologies: as was presented above (4.1.2.), we propose complex aspectual classes such as ACH•OBJET, ACC•OBJET, etc., on the one hand, and complex classes such as ACH-ETAT and ACC-ETAT on the other hand. As for partial verb-noun aspectual discrepancies, we distinguish cases where:

1. the verb belongs to a complex aspectual class whereas the nominalization belongs to a simple class, which is a subclass of the verb's complex class. A "reduction" of the verb's complex aspectual class is thus at play; this is the case for over 37% of verb-noun pairs (25 out of 67), e.g. ACCUSER (ACH-ETAT) → ACCUSATION (ACH);
2. the noun belongs to a complex aspectual class where one of the subclasses corresponds to either a simple verbal class or one of the verbal complex class constituents. An elaboration on the verbal aspectual class is thus at play; this is the case for over 62% of verb-noun pairs (42 out of 67), e.g. DÉFINIR (ACC) → DÉFINITION (ACC•OBJET).

Verbal/nominal aspect correspondence		Total
Perfect match		492 (66.8%)
Mismatch	total	67 (9.1%)
	partial	64 (8.7%)
Other		114 (98 OBJ / 16 HAB)

Table 10: Verb-noun aspectual discrepancies

Cases summed up in the last line of table **??** are mismatches stemming from a difference between verbal and nominal aspectual ontology.

5.3. Comparing both methods of aspectual class attribution

In this project, we have used two different semantic annotation methods: one based on transformation tests applied on real-life sentences by naive annotators, the other based on forged sentences applied by linguistically trained annotators. In this section, we wish to assess whether both methods yield the same classes or not.

As can be seen in table 11, the degree of correspondence between aspectual classes assigned by each method is very high: for events, 2,001 matches out of 2,309 cases; for states, 136 matches out of 217, and for objects 211 out of 232.

CA	nb occ	distribution in lexicon
EVT	**2,309**	EVT (**2,001**), STATE (94), OBJECT (153), Other (61)
STATE	**217**	STATE (**136**), EVT (53), OBJECT (22), Other (6)
OBJECT	**232**	OBJECT (**211**), EVT (19), STATE (0), Other (2)

Table 11: Comparison of semantic class attribution based on two different methods

$\sqrt{((PoxQo)/n)}$, where Po is the percentage of the observed property, Qo the complementary percentage.

[8] Standard χ^2 with 18 degrees of freedom.

The differences stem in most cases from aspectual encoding errors in the lexicon. Thirteen occurrences of lexeme PROCÉDURE (in the sense of *legal procedure*) are, for instance, labeled EVT in the corpus whereas this lexeme appears as an OBJECT in our lexicon, while this lexeme denotes an activity and thus an event. Other mistakes can be observed, such as: ADMINISTRATION#2 (in the sense of *set of persons in charge of the administration of something*) which has been described as occurrences of ADMINISTRATION#1 (*the resulting state of the process*). Confronting data extracted from our corpus and data from our lexicon thus allows us to ensure their quality.

6. Conclusion

We have presented in this paper a corpus-based semantic annotation project. The resulting annotated corpus is the groundwork for one of the main outcomes of the project: a semantic and syntactic electronic lexicon for French deverbal nouns, linked to their occurrences in the French Treebank[9]. This lexicon will be the first, so far as we know, to propose a description of aspectual properties for French nouns, in the continuity of projects such as Nomlex (Macleod et al., 1998) and SIMPLE (Bel et al., 2000).

By combining theoretical and empirical approaches to linguistic description, the Nomage project provides stable data available for further analysis regarding nominal aspect. The interaction between both approaches has proven its interest. On the one hand, theory provides the empirical approach with linguistic tests and an ontology. On the other hand, the theoretical approach is challenged by contextual data, which raise the question of vagueness and of the relevance of the theoretical classes.

The relationship between the verbal and nominal aspectual systems also has to be further investigated. There are structural differences, due to the grammatical specificities of each category, that should be questioned. For instance, as long as there are no OBJECT verbs, under which conditions do verbs yield OBJECT nominalizations? Does the mass-count nominal feature correspond to some lexical property in the verbal domain? How can the cases of conversion (e.g. MARCHE MARCHER) be analyzed with regard to aspectual inheritance?

In future work, we intend to extend the semantic annotation process to French deadjectival nouns (e.g. FIDÉLITÉ from FIDÈLE), and to non deverbal predicative nouns (e.g. crime). We also intend to extend our methodology to other languages: Spanish, English and Catalan.

7. References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks, Building and Using Parsed Corpora*. Kluwer, Dordrecht.

Lucie Barque, Richard Huyghe, Anne Jugnet, and Rafael Marín. 2009. Two types of deverbal activity nouns in French. In *5th International Conference on Generative Approaches to the Lexicon*, pages 169–175, Pisa.

Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowsky, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. In *Proceedings of LREC 2000*, pages 1379–1384, Athens.

Antonietta Bisetto and Chiara Melloni. 2008. On the Interpretation of Nominals: Towards a Result-Oriented Verb Classification. In *Proceedings of the 40th Linguistics Colloquium*, Frankfurt.

Noam Chomsky. 1970. Remarks on nominalizations. In A.J. Roderick & P.S. Rosenbaum, editor, *Readings in English Tranformational Grammar*. Ginn and Co, Waltham (MA).

Georgette Dal and Fiammetta Namer. 2010. Les noms en *-ance/-ence* du français : quel(s) patron(s) constructionnel(s)? In *Actes en ligne du 2e Congrès Mondial de Linguistique Française*, pages 893–907, La Nouvelle Orléans, États-Unis.

David Dowty. 1979. *Word Meaning and Montague Grammar*. D. Reidel Publishing Co., Dordrecht.

Nelly Flaux and Danièle Van de Velde. 2000. *Les noms en français : esquisse de classement*. Ophrys, Paris.

Livio Gaeta. 2002. *Quando i verbi compaiono come nomi. Un saggio di morfologia naturale*. FrancoAngeli, Milano.

Danièle Godard and Jacques Jayez. 1996. Types nominaux et anaphores : le cas des objets et des événements. In W. De Mulder, L. Tasmowki-De Ryck, and C. Vetters, editors, *Cahiers Chronos 1*.

Jane Grimshaw. 1990. *Argument structure*. MIT Press, Cambridge, MA.

Pauline Haas and Richard Huyghe. 2010. Les propriétés aspectuelles des noms d'activités. *Cahiers Chronos*, 21.

Pauline Haas, Richard Huyghe, and Rafael Marín. 2008. Du verbe au nom : calques et décalages aspectuels. In *Congrès Mondial de Linguistique Française (CMLF 2008)*, pages 2039–2053, Paris.

Sophie Heyd and Marie-Laurence Knittel. 2009. Les noms d'activité parmi les noms abstraits : propriétés aspectuelles, distributionnelles et interprétatives. *Lingvisticæ investigationes*, 32-1.

Robert B. Lees. 1960. *The Grammar of English Nominalizations*. Mouton, The Hague.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barret, and Ruth Reeves. 1998. NOMLEX: A Lexicon of Nominalizations. In *Proceedings of Euralex'98*, Liege, Belgium.

Igor Mel'čuk. 2004a. Actants in Semantics and Syntax I: actants in semantics. *Linguistics*, 42(1):1–66.

Igor Mel'čuk. 2004b. Actants in Semantics and Syntax II: actants in syntax. *Linguistics*, 42(2):247–291.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekeley, Veronkia Zielinska, and Brian Young. 2004. The Cross-Breeding of Dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.

Adam Meyers. 2007. Annotation Guidelines for NomBank. *Online publication:*

[9] A simple query interface to the Nomage lexicon can be accessed at `http://nomage.recherche.univ-lille3.fr/webgui`.

http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf.

Jasmina Milicévic and Alain Polguère. 2010. Ambivalence sémantique des noms de communication langagière en français. In *Congrès Mondial de Linguistique Française (CMLF 2010)*, Paris.

Aina Peris, Mariona Taulé, and Horacio Rodríguez. 2010. Semantic Annotation of Deverbal Nominalizations in the Spanish corpus AnCora. In *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, University of Tartu, Estonia.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

Ludovic Tanguy and Nabil Hathout. 2002. Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Actes de TALN 2002*, Nancy.

M. Taulé, M.A. Martí, and M. Recasens. 2008. Ancora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation*, Marrakesh, Morocco.

Karel Van den Eynde and Piet Mertens. 2003. La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104.

Evaluating Morphological Resources: a Task-Based Study for French Question Answering

Delphine Bernhard[1], **Bruno Cartoni**[2], **Delphine Tribout**[1]

(1) LIMSI-CNRS, 91403 Orsay, France
(2) Département de linguistique, Université de Genève, Suisse
bernhard@limsi.fr, bruno.cartoni@unige.ch, tribout@limsi.fr

Abstract

Morphology is a key component for many Natural Language Processing applications. In this article, we focus on one prototypical application, namely Question Answering (QA). In QA, morphological relations, especially those relying on the derivation and compounding processes, are often addressed in a superficial manner. Considering that some resources are able to provide deep and precise knowledge about a large spectrum of morphological processes, the issue lies first in determining the morphological phenomena which are most relevant for QA systems and second in evaluating the coverage of existing resources in this respect. To this aim, we describe a manual annotation and analysis of French question-answer pairs, which was performed in order to produce a unique and well-characterised reference dataset. Based on this study, we evaluate five different morphological resources for French and show that some dedicated resources are still lacking, which would cover phenomena such as denominal adjectives and agent deverbal nouns.

1. Introduction

Morphological resources are central to many Natural Language Processing applications. Despite their importance, resources are still lacking for many languages and domains, in particular with regard to constructional morphology, i.e. derivation and compounding. Moreover, they are usually evaluated intrinsically by human evaluators. As for extrinsic evaluations, they focus on the performance gains which can be obtained by using morphological knowledge in a specific applications, e.g. speech recognition (Creutz et al., 2007), machine translation (Koehn and Hoang, 2007) or information retrieval (Hahn et al., 2003). In this article, we propose a new method for evaluating resources which consists in manually building a task specific gold-standard in order to measure the coverage and quality of morphological resources. Here we focus on one prototypical application, namely Question Answering (QA).

QA systems aim at providing a precise answer to a given user question. To this aim, they usually rely on an Information Retrieval (IR) component which attempts to match words in the question and words in the text passages containing a potential answer. The major difficulty lies in the lexical gap problem, which occurs when a document is related to a question even though it does not contain the same words as the question. QA and IR systems must thus find a way of retrieving relevant documents without relying only on mere identity between words. In this context, morphology has often been preferred over semantics because the integration of morphological knowledge is easier. Research in IR and QA has thus tried to incorporate morphological knowledge, either by expanding the query, by indexing documents with morphologically motivated units or by using question reformulation or rephrasing patterns to identify the answer.

Most of the research carried out so far made use of simple heuristic-based stemming techniques which cut off word endings (such as (Lennon et al., 1988), (Harman, 1991), (Fuller and Zobel, 1998)). These turned out to be rather efficient for languages with a "less-rich" morphology, such as English, but they are not available for all languages (McNamee et al., 2009). In most cases, the recall is slightly improved, but these techniques also produce some noise, as shown by the example described in Bilotti et al. (2004): *organisation* and *organ* are stemmed to the same form by the Porter Algorithm. Another interesting piece of research, described in Moreau and Claveau (2006), shows that extending the query by morphological knowledge significantly improves the results, in most of the European languages for which they performed the experiment. To acquire morphological knowledge, they made use of a learning method based on analogy techniques. Consequently, they captured only affixation processes, and moreover only transparent affixation processes (that share a rather long character string), leaving aside conversion, reduction processes, or affixation on suppletive forms. They also admitted that, even with some precautions (long minimal character string, etc.), some incorrect pairs of morphologically related words are captured (*pondre* with *répondre*).

As we have shown, QA applications mostly rely on partial or superficial morphological knowledge. Moreover, only few studies specifically address the role of morphology within such systems. Most of the evaluations are extrinsic (based on the measurement of the improvement of an entire system when a morphological "module" is applied), and globally, the use of morphology (either indexing or query expansion) is very coarse.

However, some morphological resources are now able to provide detailed and precise knowledge about a large spectrum of morphological processes. The issue is more in weighting the relations to be implemented, and in determining the resources to be used – or built if lacking. Hence, we address two specific research questions in this article:

1. What morphological phenomena are most relevant in a QA application?
2. How well do available resources for French morphol-

ogy cover these phenomena?

These two aspects are linked together because we first need to characterize the morphological relations wich are relevant in a QA task in order to evaluate the use of existing morphological resources in a QA system.
We therefore performed our evaluation of morphological resources for French in two steps. First, we have manually annotated and analysed pairs made of a question on one side, and the snippet containing the answer on the other side, in order to determine the morphological relations involved. Secondly, we used this set of pairs of morphologically related words as a gold-standard to evaluate the coverage of available resources for French. Since the gold-standard has been carefully characterised, precise measures can be computed for different morphological processes.
The contributions of the paper are as follows:

- We present the constitution and the analysis of a unique gold-standard for morphological relations, based on a detailed annotation of three different corpora of question-answer pairs. This study provides important insights on the type of morphological knowledge to be integrated into QA systems in order to improve their performance.
- We evaluate and compare five different morphological resources for French, including both inflectional and derivational morphology.
- We show that resources covering some important morphological phenomena are still lacking for the French language and make concrete proposals about the resources which would be most helpful for QA.

2. Annotation of Question-Answer pairs

2.1. Description of the datasets

The datasets gathered for the annotation come from three very different QA corpora: Quæro, EQueR-Medical and Conique, which are presented below. Our aim in annotating different types of corpora was to determine if there are significant differences in the morphological processes observed depending on the type of data. Table 1 presents statistical information on each corpus.

Quæro The French Quæro corpus has been built for QA evaluation (Quintard et al., 2010) whithin the Quæro project. The corpus consists of 2.5M French documents extracted from the web and a set of 250 questions for the 2008 evaluation and 507 questions for the 2009 evaluation. The document corpus has been constituted by taking the first 100 pages returned by the Exalead search-engine for a set of requests found in the search-engine's logs. As for the questions, they have been written by French native speakers by using the contents of the documents for the 2008 evaluation, and by using only the query logs of the search-engine for the 2009 evaluation. There are three types of questions: factual questions, boolean questions which ask for a yes-no answer and questions requiring a list for answer.
We have constituted our corpus for the annotation task by taking all the snippets returned by the Ritel-QA System (Quintard et al., 2010) that have been manually validated as containing the correct answer for each factual question of the two evaluation campaigns. We thus obtained 566 pairs of question and snippet containing the answer, 338 from the the 2008 evaluation and 228 from the 2009 evaluation.

EQueR-Medical The EQueR evaluation dataset has been constituted within the EQueR-EVALDA evaluation campaign for French Question Answering systems (Ayache et al., 2006). The campaign included two main tasks: (i) general domain QA over a collection of newspaper articles and senate reports and (ii) specialised domain QA over a collection of medical texts. We restricted our annotation study to the medical questions. The answer snippets were retrieved by the participant systems and manually validated by a specialised judge.
Overall, the EQueR-Medical dataset comprises 394 question answer-snippet pairs for 200 different questions.

CONIQUE The CONIQUE corpus has been built with the objective of studying relevant answer justifications for QA systems (Grappy et al., 2010). Answer justifications provide additional material to the user, so that she/he may trust the answer retrieved by the system. The corpus is based on a subset of 291 questions from the French EQueR campaign (Ayache et al., 2006) and several CLEF campaigns. Candidate answer snippets have been retrieved from the French Wikipedia using a coarse retrieval mechanism and manually annotated by seven annotators. In contrast to the two previously described datasets, answer snippets in CONIQUE do not correspond to the output provided by QA systems. It therefore constitutes an almost full recall dataset, devoid of any bias inherent to QA systems such as high question-snippet token overlap.
We automatically pre-processed the annotated corpus to retrieve question-snippet pairs. We only kept full or partial justifications. Moreover, we reduced the snippet to up to three sentences, centred on an annotated justification. Overall, the dataset we annotated comprises 664 question-answer pairs, for 201 different questions.

2.2. Annotation methodology

The annotation was manually performed by three trained independent annotators,[1] using the YAWAT alignment tool (Germann, 2008). YAWAT was originally developed to align words in bilingual sentence-pairs for machine translation evaluation. In our case, we aligned words and phrases in question-answer pairs and typed their morphological relation. We defined three tags for morphological relations: one for inflection, another for derivation and another for compounding. Since there can be more than one morphological step between two morphologically related words we defined specific guidelines for the annotation.
First, we did not annotate inflectional variants of auxiliaries and determiners, as these tend to be very frequent but do not provide any interesting semantic information for use in QA. Second, derivation and compounding supersede inflection. For instance, in the QA pair presented in Figure 1 there are two morphological steps between the noun *Australie* (eng:

[1] Co-authors of the present article.

	Quæro	EQueR-Medical	CONIQUE
#Questions	350	200	201
#QA pairs	566	394	664
Avg. question length	8.8	9.9	11.4
Avg. answer length	38.5	29.0	92.4

Table 1: Annotation corpora statistics

	Inflection		Derivation		Compounding	
Corpus (qa pairs)	nbr	%	nbr	%	nbr	%
Conique (664)	159	41.8	188	**49.5**	33	8.7
Quæro (566)	136	**61.8**	80	36.4	4	1.8
EQueR (394)	69	26.4	81	31.0	111	**42.5**

Table 2: Inflection, derivation and compounding in the three corpora

Australia) in the question and the feminine adjective *australienne* (eng: australian) in the answer: the first step is the derivation of the adjective *australien* (eng: australian) out of the noun, and the second one is the inflection of the derived adjective in a feminine form. But the relevant morphological relation between the question and the answer is the derivation of the adjective *australien* out of the noun *Australie*, so that only this one has been annotated. Finally, a specific tag "other" was used to label words that are not directly related (i.e. that are related by more than one morphological process).

Q: Quelle est la capitale de l' **Australie** ?
A: le territoire sur lequel est située la capitale fédérale **australienne**, Canberra .

Figure 1: Example of QA pair where both derivational and inflectional information are available

3. Analysis of the annotated data

At the end of the annotation step, we obtained a set of morphologically related words, that can be studied according to different points of view. First we studied the repartition of morphological relation types such as inflection, derivation and compounding in the three corpora. Then, we analysed in more details the part-of-speech involved in each morphological relation, the grammatical features expressed by the inflectional processes and the semantic types of derivational processes.

3.1. Morphological relation types

	Adjectives		Nouns		Verbs	
	nbr	%	nbr	%	nbr	%
Conique (159)	45	28.3	43	27.0	71	**44.7**
Quæro (136)	9	6.6	55	40.5	72	**52.9**
EQueR (69)	22	31.9	33	**47.8**	14	20.3

Table 3: Parts of speech involved in inflectional processes

The results of the annotation of each corpus according to the different types of morphological relations are presented in Table 2. Each question-answer pair (qa pair) does not necessarily contain a morphological relation, and, more importantly, several pairs of words in the same question-answer pair can be morphologically related to one another, with different morphological relations.

The figures in Table 2 show that each corpus seems to favour a particular type of morphological relation: the Conique corpus contains a majority of derivational relations, while the Quæro corpus comprises more inflectional morphology. As for the EQueR corpus, it presents more compounding than any other kind of morphological relation. Moreover, if we consider the type of morphological relation depending on the corpus, inflection has the greatest proportion in the Quæro corpus, derivation is proportionally more present in the Conique corpus, while compounding is almost absent from Conique and Quæro and very important in EQueR.

The Conique and Quæro corpora show little difference with respect to the proportion of compounding. However, Conique contains more derivational relations than Quæro does. This is due to the way the Conique corpus has been built. It is not based on the output of a QA system, but the answers have been manually retrieved and annotated. QA systems usually have difficulties in dealing with derivational morphology. Moreover, there is a large variation in question and answer length between both corpora, as shown in Table 1. This could also explain the presence of more derivationally related pairs of words in Conique, because the longer the questions and the answers, the more opportunities to observe a derived word and its base. As for EQueR, the great proportion of compounds is certainly related to domain of the corpus: it contains a lot of medical terms, which are often compound nouns, as shown in Figure 2. These morphological characteristics of medical data have already been pointed out by Namer and Zweigenbaum (2004).

In the remainder of this section we detail the annotation results for inflection and derivation only, since there are no morphological resources devoted to compounding which could be evaluted.

	direct relation		two steps		two complex	
	nbr	%	nbr	%	nbr	%
Conique (188)	174	92.6	2	1.0	12	6.4
Quæro (80)	70	87.5	1	1.3	9	11.2
EQueR (81)	70	86.4	3	3.7	8	9.9

Table 4: Derivational steps between the word in the question and the word in the anwser

	Conique (174)		Quæro (70)		EQueR (70)	
	nbr	%	nbr	%	nbr	%
Noun > Adj	41	23.6	16	22.9	**28**	**40.0**
Proper N > Adj	**45**	**25.9**	8	11.4	1	1.4
Noun > Noun	29	16.7	5	7.1	7	10.0
Proper N > N	6	3.4	8	11.4	2	2.9
Adj > Noun	3	1.7	0	0	4	5.7
Verb > Noun	41	23.6	**30**	**42.9**	25	35.7
Other	9	5.1	3	4.3	3	4.3

Table 5: Derivational processes in question-answer pairs

Q: Quelle est la conséquence de la **corticothérapie** sur l' os?
A: Le problème essentiel des **corticoïdes** réside dans leurs effets secondaires (... ostéoporose, ostéonécrose aseptique des têtes fémorales ou parfois humérales ...).

Figure 2: Example question-answer pair from EQueR

3.2. Inflection

The analysis of the inflectional relations found in the three corpora confirms the difference, already observed at the relation type level (Section 3.1.), between Conique and Quæro on the one hand and EQueR on the other hand. Indeed, in Conique and Quæro most inflectional relations are verbal, whereas in EQueR most of them are nominal and the verbal ones are very few, as shown in Table 3.

3.3. Derivation

As shown in Table 2, derivation is important in the three corpora (between 30% and 50% of the pairs). In some cases the word in the question and the word in the answer are morphologically related by more than one derivational step. For instance *lune* (eng: "moon") and *alunissage* (eng: "landing on the moon") or *lait* (eng: "milk") and *allaitement* (eng: "breastfeeding"). In these cases one word is more complex than the other, but the complex word is not directly derived from the less complex. In some other cases, like *joueur* (eng: "player") and *jouable* (eng: "playable") the word in the question and the word in the answer are morphologically related but neither derives from the other. Instead, they both derive from another word, which is *jouer* (eng: "play") for *joueur* and *jouable*. Table 4 shows the proportion of direct derivational relations, non direct derivational relations and cases where both words are complex and derive from another word. The figures show that most derivational relations between a word in the question and a word in the answer are direct (between 86% and 92%). Only very few relations are non direct. There is little to say about the case when the derivational relation is non direct, since in that case the relation between the two words is pretty unpredictable. That is why we focus our study on the pairs which contain one base and one derivative, with only one derivational process between the two.

While focusing on the direct derivational relations, we can evaluate the proportion of different derivational processes involved. Table 5 presents the result of this evaluation. The figures in Table 5 show that the corpora differ with respect to the derivational processes used. While Conique shows more denominal adjectives (about 47% of the derivational processes), Quæro and EQueR seem to favor noun formation processes (with respectively 61% and 54% of the derivational processes). These two particular derivational processes are described below.

3.3.1. Denominal adjectives

In our data, adjectives which derive from a proper noun (Proper N) are always relational adjectives, like *chilien* (eng: "chilean") derived from *Chili* (eng: "Chile"), or *africain* (eng: "african") derived from *Afrique* (eng: "Africa"). Adjectives deriving from a common noun are mostly relational adjectives too, as shown by the figures in Table 6. For instance *présidentiel* (eng: "presidential") derived from *président* (eng: "president"), or *solaire* (eng: "solar") derived from *soleil* (eng: "sun"). However there also are some qualifying adjectives. For instance *âgé* (eng: "old") which derives from *âge* (eng: "age") with the meaning 'having a certain age' or *montagneux* (eng: "mountainous") derived from *montagne* (eng: "mountain") with the meaning 'full of mountains'. Table 6 presents the proportion of relational or qualifying adjectives in our corpora, and shows that relational adjectives are much more frequent in the three corpora. It is also worth noting that the highest proportion of relational adjectives is found in the medical corpus, which confirms previous works such as (Deléger and Cartoni, 2010).

	Relational Adj.		Qualifying Adj.	
	nbr	%	nbr	%
Conique (41)	23	56.1	18	43.9
Quæro (16)	10	62.5	6	37.5
EQueR (28)	24	85.7	4	14.3

Table 6: Types of denominal adjectives

3.3.2. Noun formation processes

As regards the noun formation processes, the three corpora favour deverbal nominalisations, but deadjectival and denominal nominalisations are also found.[2] The formations of noun out of a noun are very few, except in Conique. Those are mostly masculine and feminine profession names, like *infirmier* and *infirmière* (eng: "male/female nurse"), *directeur* and *directrice* (eng: "male/female director"), *président* and *présidente* (eng: "male/female president"), which we considered to be two distinct words rather than one and the same word inflected for gender. There are some suffixed diminutive nouns too, like *rame* (eng: "ream") > *ramette* (eng: "small ream"), and prefixed nouns like *président* (eng: "president") > *vice-président* (eng: "vice-president"). We also considered the formation of a noun out of a proper noun to be a denominal nominalisation. These derived nouns are mostly demonyms (names for the resident of a place) which derive from a location denoting proper noun, like *Colombien* (eng: "Colombian") derived from the country name *Colombie* (eng: "Colombia"). This kind of nouns is found in the Conique and the Quæro corpora, but there are only two in the EQueR corpus, which is not surprising since it is a medical corpus.

Deadjectival nouns are very few in the three corpora. None of them is found in Quæro, and there are just a few of them in the other two corpora. These deadjectival nouns are property nouns, like *toxicité* (eng: "toxicity") which derives from the adjective *toxique* (eng: "toxic"). Not surprisingly deadjectival nouns denoting a property are mostly found in the EQueR corpus. It can be explained by the fact that the medical corpus contains a lot of disease or trouble nouns (like *toxicité* or *insuffisance* "insufficiency") which often refer to the property of being in a particular state (*toxicité* ≈ 'property of being toxic', *insuffisance* ≈ 'property of being insufficient').

As for deverbal nominalisations, they are most often event nouns in the three corpora, like *débarquement* (eng: "landing") derived from the verb *débarquer* (eng: "to land"). Event denoting nouns represent almost 85% of the deverbal nouns, as shown in Table 7. However, there also are a small number of agent nouns in the Conique and the Quæro corpora, but none in the EQueR corpus. For instance *réalisateur* (eng: "director") from *réaliser* (eng: "to direct"). And there is a small set of result nouns like *produit* (eng: "product") which derives from the verb *produire* (eng: "to produce").

[2]The type of nominalisation (deverbal, deadjectival or denominal) depends on the part-of-speech category of the base (verb, adjective or noun, respectively).

3.3.3. Other derivational processes

Other derivational processes include for instance adverbs formation out of adjectives, like *complètement* (eng: "completely") derived from *complet* (eng: "complete"), or *directement* (eng: "directly") derived from *direct* (eng: "direct"). There also are some prefixed deverbal verbs like *déboucher* (eng: "unblock") out of *boucher* (eng: "block") or denominal adjectives like *international* (eng: "international") derived from *nation* (eng: "nation"). Interestingly we observed no deadjectival verb formation (like *national* "national" > *nationaliser* "nationalize") and almost no denominal verb formation. Only four denominal verbs were found in Conique, and none in the other corpora. The absence of denominal verbs could be explained by the rather unpredictable semantic relation between a noun and a derived verb. As stated by Hopper and Thompson (1984) there is an asymmetry in the lexical categories, since a nominalisation still names the event denoted by the verb, whereas a verbalization does not refer to the entity denoted by the noun, but denotes an event associated with that entity. For instance, the noun *destruction* denotes the same event as its base verb *destruct*. But in the case of a denominal verb like *hospitalize*, the verb does not refer to the object denoted by the base noun *hospital*, but denotes some event related to that object. What is more, the events we could associate to an entity are numerous and various. So, the semantic relation between a noun and its derived verb is less informative than that of a verb and its derived noun. It is not surprising then that so few nouns related to their derived verbs were found in the corpora.

4. Evaluation of morphological resources

The set of morphologically annotated data presented in the previous section forms a gold-standard of morphological relations on which we can evaluate the coverage of existing morphological resources.

4.1. Description of the resources

Several resources are available to deal with French morphology. However none of them handles the whole morphology for French. Instead, there are different resources, each of them dealing with a specific area of French morphology. Thus, we took the morphological resources dealing with each type of morphological process we found in the corpora and evaluated them according to their morphological specificity. For inflectional morphology we evaluated two resources : Morphalou and Le*fff*. For derivational morphology we evaluated three different resources : Verbaction and Dubois for deverbal nouns, and Prolexbase for denominal adjectives.

4.1.1. Morphalou

Morphalou is an inflectional lexicon for French.[3] It contains 539,413 inflected forms corresponding to 68,075 lemmas.

[3]http://www.cnrtl.fr/lexiques/morphalou/

	Conique (41)		Quæro (30)		EQueR (25)	
	nbr	%	nbr	%	nbr	%
Verb > Event N	**34**	**82.9**	**25**	**83.3**	**22**	**88**
Verb > Agent N	4	9.8	4	13.3	0	0
Verb > Other N	3	7.3	1	3.4	3	12

Table 7: Semantic types of deverbal nouns in question-answer pairs

4.1.2. Le*fff*

Le*fff* is a syntactic and morphological lexicon for French (Sagot, 2010).[4] It contains morpho-syntactic information for each inflected form, suchs as part of speech, lemma and sub-categorization. Overall it contains 534,763 infected forms.

4.1.3. Verbaction

Verbaction is a lexicon of French action nouns linked to their corresponding verbs (Hathout et al., 2002; Hathout and Tanguy, 2002,).[5] It totals 9,393 verb-noun pairs.

4.1.4. Dubois

This XML resource is based on the description of French verbs by Dubois and Dubois-Charlier (1997).[6] It classifies verbs in semantic and syntactical classes and also provides information about some derivatives of the verbs. Overall it contains 25,609 verb entries and mentions 33,955 derivatives.

4.1.5. Prolexbase

Prolexbase is a multilingual dictionary of proper nouns (Bouchou and Maurel, 2008; Tran and Maurel, 2006).[7] While not targeted at morphology, it nevertheless provides information about relational nouns and adjectives associated with proper nouns, e.g. *Français* and *français* (eng: "French") are explicitly associated with *France*. In some cases, relational nouns and adjectives are not morphologically related to the proper noun, e.g. *britannique* (eng: "british") with *Royaume-Uni* (eng: "United Kingdom"). Overall, it comprises 76,118 lemma and 20,614 derivational relations.

4.2. Evaluation results

4.2.1. Inflection

Two resources, Morphalou and Le*fff*, have been evaluated regarding the inflectional phenomena. Both resources contain approximately the same amount of inflected forms (see previous section), but have been built using different methods. Part of the information in Le*fff* has been automatically acquired and manually validated, while Morphalou's data originate from the TLFNome, the nomenclature of the TLF (Trésor de la Langue Française). In order to evaluate the coverage of the resources, each member of the inflectionally related word pairs identified in our corpus study was looked up in the lexicon. If correctly analysed, both members of the pairs should have the same lemma, and the link between them can be computed. The coverage of each resource was calculated by considering pairs that were correctly analysed, i.e. pairs of words with the same lemma. Table 8 presents the result of the evaluation of Le*fff* and Morphalou for inflectional pairs.

Both resources have very high coverage of inflectional processes in the three corpora. Le*fff* appears to be a little more complete than Morphalou, since its coverage is slightly better in the Conique corpora. Moreover, on the EQueR dataset they differ in the word pairs they cover although they have the same global coverage. Indeed, both of them cover 65 pairs of inflected words out of 69, but the covered pairs are not exactly the same. So that the global coverage made by at least one resource is slightly better than the coverage of one and only one resource. This fact shows that using two different resources for the same type of morphological phenomena can improve the global coverage of the data.

4.2.2. Derivation

Assessing derivational resources is not as straightforward as inflectional ones. The three considered morphological resources that are available for French derivational morphology are designed to address specific morphological phenomena. Dubois and VerbAction contain exclusively deverbal morphology, while Prolexbase only contains demonym nouns and relational adjectives. Consequently, assessing the relevance of these resources can only be done with the appropriate sub-part of the gold-standard. The coverage of VerbAction and Prolexbase was calculated by counting the number of pairs that have been found in them. As for Dubois, it does not literally contain verbal derivatives. Those are only mentioned with specific information from which we can deduce the derivatives. Thus, in order to evaluate the coverage of Dubois we only took into account cases where the derivatives would be automatically computable from information provided in the resource.

As regards the deverbal nouns, Table 9 summarises the coverage of VerbAction and Dubois for event nouns. As we can see, VerbAction has a better coverage than Dubois, especially in lay corpora (Conique and Quæro). As for the deverbal agentive nouns, Dubois covers 100% of the Conique corpus and 75% of the Quæro corpus (no agentive noun has been found in EQueR corpus), while VerbAction does not contain any of them, since it is devoted to action nouns.

As for the demonyms and relational adjectives derived from geographical names, the result of the evaluation of Pro-

[4] http://alpage.inria.fr/~sagot/lefff.html

[5] http://redac.univ-tlse2.fr/lexiques/verbaction.html

[6] http://rali.iro.umontreal.ca/Dubois/

[7] http://www.cnrtl.fr/lexiques/prolex/

Corpus (nbr.)	Lefff		Morphalou		Global coverage	
	nbr.	%	nbr.	%	nbr.	%
Conique (159)	159	100.0	157	98.7	159	100.0
Quæro (136)	135	99.3	135	99.3	135	99.3
EQueR (69)	65	94.2	65	94.2	66	95.6
Total (364)	**359**	**98.6**	**357**	**98.1**	**360**	**98.9**

Table 8: Coverage of inflection

Corpus (nbr.)	VerbAction		Dubois	
	nbr.	%	nbr.	%
Conique (34)	33	97.1	19	55.9
Quæro (25)	25	100.0	9	36.0
EQueR (22)	22	100.0	19	86.4
Total (81)	**80**	**98.8**	**47**	**58.0**

Table 9: Coverage of resources for deverbal event nouns

lexbase is presented in Table 10. We distinguished between Demonym, Relational adjective, and LocOrg (grouping name of place and institutional entities). The figures show that Prolexbase has a very good coverage for both Demonyms derived from a Location name, and relational adjectives derived from Demonyms or Location names. In the Quæro corpus no Demonym>RelAdj pair has been found, and in the EQueR corpus, only one pair LocOrg>RelAdj has been found and is correctly analysed in Prolexbase.

When evaluating the three different types of derivational resources (VerbAction, Dubois and Prolexbase) on the whole gold-standard the coverage is not as high as on specific parts of the gold-standard. Indeed, the global coverage of the three resources is only slightly higher than 50%, as shown in Table 11. Morphological resources are efficient for specific morphological processes. But very frequent phenomena seem to be lacking in the assessed resources, like deverbal agent nouns formation[8] and denominal adjectives formation. This is highly regrettable since the latter process is the second most frequent in the pairs in Conique and Quæro, and the first more frequent in EQueR, as shown in section 3. Consequently, efforts on building resources should be put on this particular phenomenon to address such a frequent issue.

5. Conclusion and Perspectives

In this paper, we have presented an in-depth analysis of the role of morphology in one specific NLP task: Question Answering. Based on a large-scale annotation of three distinct corpora of question-answer pairs, we have built a gold-standard of morphologically related words in question-answer pairs. This gold-standard provides interesting insights on the kind of morphological relations that are mostly implied, and it uncovers those which could have a significant impact on the application performance. Moreover, based on this gold-standard, we have evaluated the coverage of existing morphological resources for French. This evaluation proved that the analyzed French inflectional and derivational resources have a good coverage of the morphological knowledge they target. But some important morphological phenomena are lacking a dedicated resource such as denominal adjectives and agent deverbal nouns. In the future, we hence plan to develop some new French morphological resources for these two phenomena.

[8]Dubois does contain information about deverbal agent nouns. However, these nouns are not explicitly part of the resource and would have to be automatically computed from the indications provided in the resource.

Acknowledgments

This work has been partially financed by OSEO under the Quaero program.

6. References

Christelle Ayache, Brigitte Grau, and Anne Vilnat. 2006. EQueR: the French Evaluation campaign of Question-Answering Systems. In *Proceedings of LREC 2006*.

Matthew W. Bilotti, Boris Katz, and Jimmy Lin. 2004. What Works Better for Question Answering: Stemming or Morphological Query Expansion. In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, Sheffield, England.

Béatrice Bouchou and Denis Maurel. 2008. Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres. *Traitement Automatique des Langues*, 49(1):61–88.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1):1–29.

Louise Deléger and Bruno Cartoni. 2010. Adjectifs relationnels et langue de spécialité : vérification d'une hypothèse linguistique en corpus comparable médical. In *Proceedings of TALN 2010*.

Jean Dubois and Françoise Dubois-Charlier. 1997. *Les verbes français*. Larousse-Bordas.

Corpus	Morphological relation (nbr.)		Found in Prolexbase	
			nbr.	%
Conique	Demonym - Rel Adj	(1)	1	100.0
	LocOrg - Demonym	(6)	6	100.0
	LocOrg - Rel Adj	(45)	43	95.6
Quæro	LocOrg - Demonym	(8)	5	62.5
	LocOrg - Rel Adj	(8)	8	100.0
EQueR	LocOrg - Rel Adj	(1)	1	100.0
Total		**69**	**64**	**92.7**

Table 10: Coverage of Prolexbase for Geographic morphological relation

Corpus (nbr.)	Global coverage	
	nbr.	%
Conique (174)	98	56.3
Quæro (70)	41	58.6
EQueR (70)	26	37.1
Total (314)	**165**	**52.5**

Table 11: Global coverage of the three derivational resources on derivational pairs

Michael Fuller and Justin Zobel. 1998. Conflation-based comparison of stemming algorithms. In *Proceedings of the Third Australian Document Computing Symposium*, pages 8–13, Sydney.

Ulrich Germann. 2008. Yawat: yet another word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT '08)*, pages 20–23.

Arnaud Grappy, Brigitte Grau, Olivier Ferret, Cyril Grouin, Véronique Moriceau, Isabelle Robba, Xavier Tannier, Anne Vilnat, and Vincent Barbier. 2010. A Corpus for Studying Full Answer Justification. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Udo Hahn, Martin Honeck, and Stefan Shulz. 2003. Subword-Based Text Retrieval. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, Big Island, Hawaii, January 06 - 09.

Donna Harman. 1991. How effective is suffixing? *Journal of the American Society of Information Science*, 42(1):7–15.

Nabil Hathout and Ludovic Tanguy. 2002. Webaffix: Discovering Morphological Links on the WWW. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1799–1804, Las Palmas de Gran Canaria, Espagne. ELRA.

Nabil Hathout, Fiammetta Namer, and Georgette Dal, 2002. *Many Morphologies*, chapter An Experimental Constructional Database : The MorTAL Project, pages 178–209. Cascadilla Press.

P.J. Hopper and S.A. Thompson. 1984. The discourse basis for lexical categories in universal grammar. *Language*, 60:703–752.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of EMNLP-CoNLL 2007*, pages 868–876, Prague, Czech Republic.

Martin Lennon, David S. Pierce, Brian D. Tarry, and Peter Willett. 1988. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, 3(4):177–183.

Paul McNamee, Charles Nicholas, and James Mayfield. 2009. Addressing morphological variation in alphabetic languages. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 75–82, New York, NY, USA. ACM.

Fabienne Moreau and Vincent Claveau. 2006. Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, pages 181–192.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for french medical terminology: contribution of morphosemantics. *Eleventh MEDINFO International Conference*, pages 535–539.

Ludovic Quintard, Olivier Galibert, Gilles Adda, Brigitte Grau, Dominique Laurent, Véronique Moriceau, Sophie Rosset, Xavier Tannier, and Anne Vilnat. 2010. Question Answering on Web Data: The QA Evaluation in Quæro. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Benoît Sagot. 2010. The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Mickaël Tran and Denis Maurel. 2006. Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, 47(1):115–139.

A lexicon for processing archaic language: the case of XIXth century Slovene

Tomaž Erjavec[1], Christoph Ringlstetter[2], Maja Žorga[3], Annette Gotscharek[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
[2] Centre for Language and Information Processing, University of Munich
Schellingstrasse 10, 80799 Munich
kristof@cis.uni-muenchen.de, annette@cis.uni-muenchen.de
[3]maja.zorga@gmail.com

Abstract

The paper presents a lexicon to support computational processing of historical Slovene texts. Historical Slovene texts are being increasingly digitised and made available on the internet but are still underutilised as no language technology support is offered for their processing. Appropriate tools and resources would enable full-text searching with modern-day lemmas, modernisation of archaic language to make it more accessible to today's readers, and automatic OCR correction. We discuss the lexicon needed to support tokenisation, modernisation, lemmatisation and part-of-speech tagging of historical texts. The process of lexicon acquisition relies on a proof-read corpus, a large lexicon of contemporary Slovene, and tools to map historical forms to their contemporary equivalents via a set of rewrite rules, and to provide an editing environment for lexicon construction. The lexicon, currently work in progress, will be made publicly available; it should help not only in making digital libraries more accessible but also provide a quantitative basis for linguistic explorations of historical Slovene texts and a prototype electronic dictionary of archaic Slovene.

1. Introduction

A large number of Slovene books and periodicals from the XIXth century and earlier are being made available on the internet, e.g. via the dLib.si digital library (Krstulović and Šetinc, 2005), the Slovene literary classics project at WikiSource and Google Books.[1] Human language technology support could bring increased functionality to such digital libraries, esp. for full-text search and information retrieval. The most obvious task is automatic lemmatisation of text, which abstracts away from the morphological variation encountered in heavily inflecting languages, such as Slovene. The user can thus query for e.g. *mati* (*mother*) and receive portions of text containing this word in any of its inflected forms (*matere, materi, materjo*, etc.). Support for lemmatisation, as well as morphosyntactic tagging is well-advanced for modern-day Slovene (Erjavec & Krek, 2008). However, the situation is very different for historical Slovene, where no such research has yet been carried out for the language.

Historical Slovene[2] brings with it a number of problems related to automatic processing:

- due to the low print quality, optical character recognition (OCR) produces much worse results than for modern-day texts; currently, such texts must be hand-corrected to arrive at acceptable quality levels;
- full-text search is difficult, as the texts are not lemmatised and use different orthographic conventions with different archaic spellings, typically not familiar to the user;
- comprehension of the texts for most users can also be problematic, esp. with texts older than 1850 which use the Bohoričica alphabet.[3]

We are currently developing a tool-chain for processing archaic Slovene texts which should alleviate some of these problems. The tool, called ToTrTaLe, is an extension of the ToTaLe tool (Erjavec et al., 2005), which performs tokenisation, tagging and lemmatisation, but extended with a transcription module: after tokenisation, the word-forms are first modernised as regards spelling, and only then passed on to the tagging and lemmatisation modules. This approach follows Rayson et al. (2007) in being able to use the well-developed tagging (and lemmatisation) models for contemporary language rather than having to first develop such models for historical language – a very lengthy and expensive process.[4] The approach has the further benefit of offering the contemporary words paired with archaic ones.

This paper focuses on the transcription aspect of this process which crucially depends on a lexicon or, rather, a series of lexica for the language. In previous work (Erjavec et al., 2010) we concentrated on the first steps

[1] Hladnik (2009) gives a good overview of digitisation efforts and availability of Slovene texts on the internet.

[2] In this paper we concentrate on the Slovene from the XIXth century; the problems are, of course, worse going further back in time, but even here, due to the late development of the written Slovene word and its spelling standardisation, there are substantial differences to contemporary Slovene.

[3] The Bohoričica alphabet had different conventions in writing various Slovene sounds, e.g. »*shaloft*« is the modern-day »*žalost*«, which makes it confusing for today's readers. Of course, there are also substantial vocabulary as well as syntactic differences, to contemporary Slovene.

[4] For example, annotating for lemma and morphosyntactic description 300,000 words of contemporary Slovene (Erjavec et al., 2010) took about 1,500 hours of annotator time.

(tools and work-flow) involved in manually producing a lexicon of historical Slovene. In this paper we report on the already developed lexica as used in the context of ToTrTaLe.

The rest of the paper is structured as follows: Section 2 details the process of transcription, Section 3 describes the corpora we use in our work, Section 4 the lexica that are used and being produced, Section 5 the silver-standard lexicon, to be made publicly available, Section 6 an experiment studying the current coverage of ToTrTaLe and Section 7 gives some conclusions and directions for further work.

2. Transcription

In this section we explain how modern-day equivalents are found for words in the historical texts, as this represents the main difference to processing modern-day language. The process relies on three resources:

1. A lexicon of modern-day word-forms with associated lemmas and morphosyntactic descriptions.
2. A lexicon of archaic word-forms, with associated modern-day equivalent word-form(s)[5].
3. A set of transcription patterns, giving mappings for changes in alphabets (transliteration) and common spelling changes.

In processing historical texts, the word-forms are first normalised, i.e. de-capitalised and diacritic marks over vowels removed; the latter is most likely Slovene specific, as modern-day Slovene, unlike the language of the 19th century, does not use vowel diacritics.

The following filtering steps are performed on the normalised word-form: if the normalised word-form is an entry of the archaic lexicon, the equivalent modern-day word-form has also been identified; if not, it is checked against the modern-day lexicon. Obviously, if the normalised word-form is found in the modern-day lexicon, its modern-day equivalent has been ipso-facto found as well. This order of searching the dictionaries is important, as the modern lexicon can contain word-forms which have an incorrect meaning in the context of historical texts, so the historical lexicon also serves to block such meanings. For example, the auxiliary verb form *sem* used to be written as *sim* – but in the modern lexicon this is identified as a noun, i.e. the SIM card of a mobile telephone.

If neither lexicon contains the word, the transcription patterns are tried. Many historical spelling variants can be traced back to a set of rewrite rules or "patterns" that locally explain the difference between the contemporary and the historical spelling. For Slovene, e.g., a very prominent pattern is *r→er* as exemplified by the pair *brž→beržˇ*, where the left side represents the modern and the right the historical spelling. Patterns can also be sensitive to the word boundary, as some spelling changes occur only at the start or the end of the word, e.g. *žganjem→žganjam*, where the inflectional ending *-am* has changed into modern-day *-em*. To enable this functionality the appropriate patterns make use of the special symbol, "@", e.g. *em@→am@*.

By corpus inspection we have currently developed a set of about 100 such patterns. These patterns are operationalized by the finite-state tool Vaam (Variant aware approximate matching). Vaam (Reffle, 2011) takes as input a historical word-form, the set of patters, and a modern-day lexicon and efficiently returns the modern-day word-forms that can be computed from the archaic one by applying one or more patterns; the output list is ranked, preferring candidates where a small number of pattern applications is needed for the rewrite operation. Vaam also supports approximate matching based on edit distance, useful for identifying (and correcting) OCR errors; we have, however, not yet made use of this functionality.

It should be noted that the above process of transcription is non-deterministic. While this rarely happens in practice, the historical word-form can have several modern-day equivalents. More importantly, the Vaam module will typically return several possible alternative modernisations. We currently determine the "best" transcription by choosing the most frequent contemporary word between the possible modernisations, but more advanced models are possible, which postpone the decision of the best candidate until the tagging and lemmatisation has been performed.

3. Corpora for lexicon building

To support our work on lexicon acquisition, we use several corpora of Slovene; this section gives the details of the corpora and briefly describes the concordancer used for their inspection.

3.1. Modern language corpora

For lexicon construction, including comparative studies of historical language as opposed to modern language, contemporary corpora are needed. For this purpose we are using several corpora, all based on the FidaPLUS[6] reference corpus of modern Slovene (Arhar and Gorjanc, 2007). FidaPLUS contains 600 million words, where the words have been automatically annotated with morphosyntactic tags and lemmas. The corpora we are using are the following, with the first two having been developed in the JOS[7] project (Erjavec et al., 2010):

- jos100k is a 100,000 word sampled corpus of modern Slovene, with carefully hand-validated word-level morphosyntactic and lemma annotations
- jos1M is ten times larger than jos100k but has only partially hand-validated annotations
- fpj100M is a 100 million sample from FidaPLUS, and has only automatically assigned annotations.

[5] The two lexica have in fact a somewhat more complicated structure, which is further addressed in Section 4.

[6] http://www.fidaplus.net/

[7] http://nl.ijs.si/jos/

These three corpora thus enable studying lexical phenomena choosing either very accurate annotations, but small dataset, or vice-versa. Which option is best depends to a high degree on the frequency of the phenomenon (lexica item) being inspected.

3.2. Historical language corpora

The corpus of historical language we have been mostly using so far was compiled in the scope of the project *Deutsch-slowenische / kroatische Übersetzung* 1848–1918 (Prunč, 2007). The project addressed the linguistic study of Slovene and Croatian books translated from German in the period 1848–1918, where a large portion of the effort went towards building a digital library (compiling a corpus) of the Slovene translations. To this end, the books were first scanned and OCRed, and then, for a portion of the corpus, the transcription was hand-corrected, marked-up with structural information, and, for a few books, lemmatised; this process was supported by a web interface (Erjavec, 2007).

The sub-corpus chosen for building the historical lexicon includes all the AHLib proof-read books written before the year 1900, where the oldest one was published in 1847. There are all together 71 such books, of which the majority (56) are fiction (mostly novels) while 15 are non-fiction (from self-help books for farmers, to text-books on astronomy, chemistry, etc.). All together the corpus contains approximately 2.2 million running words. While certainly small compared to most corpora of contemporary language, it is large and varied enough to have enabled us to start building the historical lexicon.

Recently, we have also collected the older materials available from the WikiSource Slovene literary classics project,[8] led by Prof. Miran Hladnik from the Ljubljana University. In the scope of this on-going project, the raw OCR of books and other materials is being hand-corrected by students. We have downloaded the currently finished transcriptions and turned them into a uniformly encoded corpus. Due to the lack of conventions in structuring Wiki entries, the quality of the automatically acquired meta-data is not very high, however, the corpus makes up for this lack by its size: our current WikiSource corpus contains over 500 publications with over 8 million words. This corpus contains, in general, more recent texts than AHLib, most from the late 19th and early 20th century.

Further historical materials are currently also being hand-corrected, which are meant to extend the scope of the corpus, currently still lacking materials from the 18th century, further into the past.

3.3. The concordancer

All the collected historical corpora are being processed by the (current version) of the ToTrTaLe tool and are then, together with the three corpora of contemporary language, made available via a dedicated Web corpus query interface, with CWB (Christ, 1994) as the backend. The concordancer enables searching and viewing the tokens, their normalised and modernised form, the used transcription pattern, and their computed morpho-syntactic description (i.e. fine-grained PoS tag) and lemma, where the view can be either Keyword in Context (KWIC) or a frequency list. The concordancer has proved to be very helpful in determining the status and preferred annotation of the historical lexical items.

4. Types of lexica

This section gives the various types of lexica used by the program, namely: lexicon of contemporary language; historical word-forms with transcriptions into contemporary language equivalents; historical words without contemporary equivalents; words missing in the contemporary language lexicon; abbreviations; and words which need to be re-tokenised in the modernisation step.

4.1. Contemporary language

The lexicon of contemporary Slovene used was extracted from the FidaPLUS corpus, where each word was automatically annotated with its morphosyntactic description (MSD) and lemma. The MSDs are compact strings that represent the morphosyntactic features of the word form, and can be decomposed into features, e.g. the MSD `Ncms` is equivalent to `Noun, Type = common, Gender = masculine, Number = singular.`

The lexicon was gathered from the corpus by extracting all the triplets consisting of the word-form, lemma and MSD. The word-forms were lowercased. Using regular expressions, entries with anomalous "words" were removed, and only those lexical items with a frequency greater than 4 were retained. With this we arrived at a lexicon, which contains about 600,000 word-forms and 200,000 lemmas.

The lexicon is large enough to cover the majority of contemporary lexis found in historical texts, i.e. it has good recall – however, its precision is relatively low, as it contains many false friends. One example (*sim)* was already mentioned; another case is *serca,* an archaic form for *srca* (*heart$_{[sg,gen]}$*), with the lemma *srce.* This form exists in the modern lexicon, but with the lemma *serec* (*horse of a grey colour*). Such word-forms have to be added to the historical lexicon, with the correct interpretation, in order to block them being retrieved from the modern lexicon.

4.2. Historical to contemporary transcriptions

The second lexicon being developed is that of manually verified historical word-forms. The approach is corpus driven, so far using the AHLib corpus, and relaying on LeXtractor (Gotscharek et al., 2010), a specialised editor for historical lexica.

[8] http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika

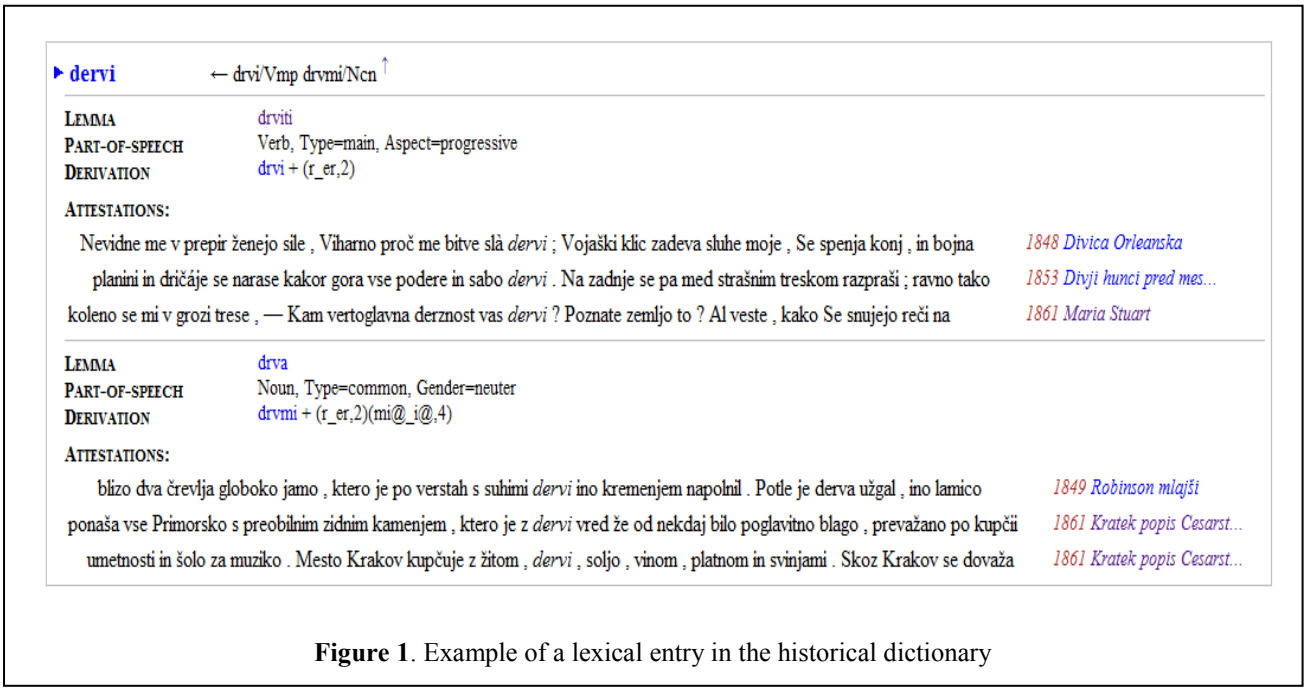


Figure 1. Example of a lexical entry in the historical dictionary

LeXtractor incorporates the Vaam pattern matching functionality and supports both a frequency based selection of entries to be added to the lexicon, as well as directly annotating word tokens in corpora. As mentioned, we give details of the manual lexicon building procedure, as well as how LeXtractor was adapted for Slovene, in Erjavec et al. (2010). Here, we will concentrate on the current structure and content of the lexica.

The lexicon we are developing has a simple structure, where each entry contains the following fields:

1. a word-form that has been witnessed in a proof-read historical text
2. the equivalent word-form from contemporary Slovene, possibly together with the patterns which map the former into the latter
3. the contemporary lemma of the word-form
4. the lexical morphosyntactic properties of the lemma
5. attestations of the word-form in the historical corpus

Figure 1 gives an example of such a lexical entry – the entry is formatted in HTML for the ease of illustration. Note that the historical word-form is ambiguous, i.e. it has two possible modern interpretations.

The intention of this manually collected lexicon is to contain the most frequently occurring archaic words in the texts; we have therefore applied frequency selection of the entries, so that closed class words are extensively covered, as are the most common open class words. We are also including as many as possible of short historical words (up to 5 characters in length) as these most frequently have false friends in the modern lexicon, either directly or via pattern application, as is the case of *sim* and *serca*.

4.3. Words without descendants

The other type of historical lexicon concerns word-forms that are missing a modern-day descendant, i.e. they do not have a corresponding contemporary lemma. For such words, LeXtractor does not currently have the functionality to enter a structured entry, apart from a comment and the attestations. Since we decided it useful to further analyse such entries, we currently enter in the comment space the following information:

1. historical lemma, as it would be written today
2. the closest contemporary Slovene synonym(s)
3. the PoS of the historical lemma
4. the source (dictionary, corpus) on the basis of which the synonyms were chosen
5. potential comments

The reasons that we are adding this information are twofold. First, by providing the "virtual" modern word-form, we are increasing the possibility of a user finding this word, even though unsure about its archaic spelling; similarly, the tagger has a greater chance of assigning the correct MSD to such a word. Secondly, while the lexicon of transcribed words is necessary for computational processing of historical texts, it is, in general, not very interesting for humans, esp. the pattern derived entries. But the words without descendants are exactly those that the modern-day reader will most likely not understand at all. So, as long as they have been identified, it is worthwhile assigning them their near-synonyms and giving the source where further information about them can be found. Such a lexicon could then also represent a prototype "bilingual" historical to modern dictionary, which is still lacking for Slovene.

4.4. Missing contemporary words

In order to improve the functionality of the tool and the filter cascade, the maintenance of the modern lexicon is crucial. Rather than modifying this lexicon directly, we, as discussed, either block inappropriate modern words by including them in the historical lexicon, or add missing words via a special lexicon. Of course, there will always

be words missing from the lexicon, and it is not our intention to add all possible contemporary words that could appear in historical texts, esp. as both the tagger and lemmatiser are able to handle unknown words. However, certain words have a rather unpredictable morphology, which causes either the tagger or lemmatiser to misinterpret them – when such cases are noticed they are added to the lexicon of missing contemporary words.

Rather than adding word-forms individually, we have implemented a Web application that is able to generate the complete inflectional paradigm given the lemma and part-of-speech. Constructing exact paradigms on the basis of this information is, in the general case, not possible, so the intention is for the lexicographer to automatically construct such a paradigm, and then edit by hand the erroneous word-forms.

4.5. Abbreviations

A lexicon very important for correct tokenisation and sentence segmentation is that of abbreviations. The tokenisation module of ToTrTaLe takes a list of abbreviations, i.e. strings ending with a full-stop, which, however do not (necessarily) end a sentence; furthermore, the period should be taken as a part of the abbreviation token. Historical language uses some abbreviations not present anymore in contemporary language – these are included to the lexicon of historical abbreviations, and then added to the tokeniser resource file. The lexicon also includes for each abbreviation its expanded form(s), although these are not currently used by the program.

4.6. Token translations

There is a final type of satellite lexicon that we use in ToTrTaLe, which is interesting from a computational perspective. In historical Slovene certain words or morphemes were written apart or together, where it is now the other way around. The most prevalent and productive example is the prefix that forms the superlative degree of adjectives: what used to be written *nar boljši* is now *najboljši*. As (word) tokens in text processing represent the basic division of characters into linguistic units, which are then further annotated, having a mismatch between archaic and contemporary Slovene at this level of description is difficult to process and encode; from being a string transcription and classification problem, the mapping of old to new language becomes one of machine translation. This is an interesting problem, esp. as it is by no means confined to historical language varieties; the same phenomenon can be found in contemporary Slovene (and other languages) where, in informal or "badly written" language people often write certain words apart, or run separate words together.

Luckily, in historical Slovene, apart from the superlative prefix, and a few other minor cases, only a well-defined set of closed-class words have changed their tokenisation. The tokeniser used by ToTrTaLe uses various classes of special lexica; one of these covers compounds, and the other "clitics", i.e. where a prefix or suffix should be split from the word, such as *-lo* in Italian. We have identified all (or most) of the closed class compounds and splits, and have also taken all the superlative adjectives found in the AHLib corpus into the compounds list. At least for these latter, this is only a stop-gap measure; in the case of Slovene superlatives, a simple regular expression (*nar .**) would cover almost all situations; as mentioned, the general case is, however, much more complicated.

The tokenisation lexicon thus contains two types of tokens, those that should be kept as one token (about 400), and those that should be split (10); in processing, these tokens are given special flags, which are retained in the output. Vaam patterns are also needed to modernise such cases, e.g. *@naj→@nar_*, where the underscore represents the space character.

entries	77,783
words	63,447
lemmas	18,940
modern entries	73,736
historical	3,181
no descendant	529
blocked modern	230
abbreviations	63
merged	44

Table 1. The size of the silver standard historical lexicon

5. Silver standard lexicon

From the partial and heterogeneous lexica we created a "silver standard" historical lexicon, which, in addition to the hand-gathered lexica also contains automatically collected "safe" modern words attested in the historical corpus. The AHLib corpus was annotated with ToTrTaLe, and the lemmas of all the contemporary words were verified against a lexicon composed of the lexicon derived from the jos100k corpus and the large Slovene monolingual dictionary SSKJ. If the automatically assigned annotations matched those in this lexicon then the entry was included in the silver standard lexicon. This approach yields highly reliable lexical entries.

Table 1 gives the size of the current lexicon, where an entry is taken to be the 4-tuple (normalised word-form, modern word-form, modern-lemma, PoS/MSD). The main part of the lexicon is contributed by modern words, while the manually collected part of historical forms currently has about 4,000 entries.

The silver standard lexicon is encoded against a slightly enhanced schema of the LeXtractor lexicon dump XML. As illustrated in Figure 2, each entry is given a type, and is headed by the (normalised) word-form. The entry can have several analyses, each giving the modernised form, lemma, PoS, possibly modern near synonyms and attestations. Entries for the tokenisation lexicon are recognised by having a white-space in the word-form or modern derivation.

```
<entry type="no_descendant">
  <wordform>alipak</wordform>
  <note>kontekst</note>
  <analyses>
    <analysis>
      <lemma>alipak</lemma>
      <pos>C</pos>
      <derivations>
        <derivation>ali pak</derivation>
      </derivations>
      <synonyms>
        <synonym>ali</synonym>
        <synonym>ali pa</synonym>
      </synonyms>
      <attestations>
        <attestation
            src="korpora/FPG06523.txt"
           position="58207">
          <pre>— ali na suhi zemlji gdè v
                Ameriki ,</pre>
          <word>alipak</word>
          <post>le na kterem ostrovi , še
                 dozdaj ni vedel</post>
        </attestation>
...
```

Figure 2 XML encoding of the lexicon.

6. Lexicon coverage

We performed an experiment in which we evaluate the coverage of the ToTrTaLe given the current lexicon(s) and pattern set. As AHLib served as the development data-set, we took for the experiment the Wiki corpus and, as the modern-day baseline, the Slovene part of the SPOOK parallel corpus of recently translated novels.[9] Both corpora were annotated with ToTaLe, and the Wiki corpus also with ToTrTaLe. We were interested in how the annotations of the two corpora differ when processed with the same model, and how the historical corpus annotations differ when processed without or with the transcription.

Table 2 gives the number and proportions of annotation classes, depending on the corpus and mode of processing. The first row gives the number of word tokens and (normalised) word types. The number of types, i.e. the size of the lexicon needed to completely cover the corpora, is quite high, but it of course also includes all the typos etc. from the source corpora.

The second row shows how many words were found in the modern FidaPLUS lexicon. The percentage is significantly lower with the Wiki corpus, esp. if we compare the number of types; from 83% with modern text down to 54% with the historical one. The third line gives the number of modern words found in the silver-standard dictionary derived lexicon; again, the number of types drops from 83% to 54%. Comparing the "Modern" and "Dictionary" number of the Wiki corpus processed with and without transcription, we note that the numbers obtained with transcriptions are slightly lower; the reason is that some words from the modern lexicon are, when using transcription, blocked by the historical lexicon.

The next line gives the number of unknown words; if Spook has about 16% unknown word types, Wiki without transcriptions has over 45%. With transcription this number drops to 39%, i.e. while we do experience some gain, we are still far from reaching modern-day recognition rates. The decrease of unknown words when using transcription can be mostly attributed to the use of patterns; they help in recognising almost 6% of word types, which is, however, only 0.5% of word tokens; and even here we have to take into account that there is no guarantee that the found modern word is in fact the correct one. The rest of the decrease in unknown words is due to the lexicon of historical words. Out of about 4,000 entries currently in the historical lexicon 2,200 were used; this is under 1% of the lexical types, i.e. much less than covered by the patterns, but, conversely, the number of tokens covered by the historical lexicon (0.76%) is greater than that covered by the patterns (0.49%).

Spook	*Tokens*	*%*	*Types*	*%*
Words	1,825,692	100.00	120,723	100.00
Modern	1,773,019	97.11	100,954	83.62
Dictionary	1,708,764	93.60	85,852	71.11
Unknown	52,673	2.89	19,769	16.38
No lemma	920	0.05	584	0.48
Wiki without transcription				
Words	8,219,093	100.00	249,262	100.00
Modern	7,868,823	95.74	135,490	54.36
Dictionary	7,522,562	91.53	109,549	43.95
Unknown	350,270	4.26	113,772	45.64
No lemma	15,796	0.19	5,623	2.26
Wiki with transcription				
Modern	7,858,325	95.61	135,490	54.36
Dictionary	7,512,988	91.41	109,550	43.95
Historical	62,822	0.76	2,231	0.90
Pattern	39,902	0.49	14,560	5.84
Unknown	258,044	3.14	97,732	39.21
No lemma	9,767	0.12	4,398	1.76

Table 2. Coverage of lexica over modern-day SPOOK corpus and 19th century Wiki corpus with and without transcription.

[9] This parallel corpus is being developed in the scope of the SPOOK project, http://lojze.lugos.si/spook/

The last line in all three tables gives that number of words that could not be lemmatised. These words are interesting, as they point to the morphological changes that occurred over time; in the modern corpus there are only 0.5% of such word types, while the Wiki without transcription has 5 times more, 2.26%; transcription lowers this number to 1.76%. Such words which cannot be lemmatised with the model for modern Slovene are very consistently true archaic words, i.e. good candidates for inclusion into the historical lexicon.

7. Conclusions

The paper presented our methodology of building a lexicon to help process historical language, in particular the Slovene of the XIXth century in the context of the ToTrTaLe tool. The background resources of this work are a historical corpus, a contemporary lexicon of Slovene, spelling variation patterns, and the Vaam and LeXtractor software.

In further work we plan to significantly enlarge the historical lexicon; now that the tools have been set-up and we have elaborated the methodology of the lexicographical work, we will engage more people to work on the lexicon, with the target size between 10 and 20 thousand entries. We plan to move from the frequency based word selection to annotating corpus tokens directly – this work also connects to our intention of compiling a gold-standard historical corpus with hand validated annotations. Such a corpus is useful for evaluating the precision/recall of various computational annotation methods and underlying resources, say the transcription rules and, of course, the lexicon. As mentioned, we will also extend the corpus with new materials, esp. newspapers and older books.

Current work has also been exclusively empirically driven, i.e. we addressed only issues that directly arose out of the lexical items found in the corpus. In the future we plan to take into account the linguistic research on historical Slovene that has been done so far , as discussed e.g. in Orožen (1996). Hopefully, our computational approach might also reveal new quantitative and qualitative linguistic insights into the language as used in XIXth century Slovenia.

The concordancer to the corpora is already publicly available at http://nl2.ijs.si/ahlib.html. We will also make the produced corpus and lexicon available under a Creative Commons licence, in the hope that it will facilitate further studies of Slovene historical language.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions. All errors in the paper of course remain our own. The work presented in this paper was supported by the EU FP7 ICT project IMPACT, “Improving Access to Text”.

References

Oliver Christ, 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. Proceedings of COMPLEX ‘94: 3rd Conference on Computational Lexicography and Text Research. 23-32, Budapest, Hungary.

Špela Arhar and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. [Corpus FidaPLUS: a new generation of the Slovene reference corpus] *Jezik in slovstvo*, 52(2).

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, and Ralf Steinberger. Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. In Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland. 2005, pp. 32-36.

Tomaž Erjavec, Simon Krek, 2008. The JOS morphosyntactically tagged corpus of Slovene. In the Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC’08, Paris. ELRA.

Tomaž Erjavec, Christoph Ringlstetter, Maja Žorga, and Annette Gotscharek, 2010. Towards a Lexicon of XIXth Century Slovene. In Proceedings of the Seventh Language Technologies Conference, October 14th-15th, 2010, Ljubljana, Slovenia. Jožef Stefan Institute.

Annette Gotscharek, Ulrich Reffle, Christoph Ringlstetter, Klaus U. Schulz, and Andreas Neumann. Towards information retrieval on historical document collections: the role of matching procedures and special lexica. International Journal on Document Analysis and Recognition, pp. 1-13, 2010.

Miran Hladnik. 2009. Infrastruktura slovenistične literarne vede [Infrastructure of Slovene Literary Studies]. In *Obdobja 28 – Infrastruktura slovenščine in slovenistike*. pp. 161–69.

Zoran Krstulović and Lenart Šetinc. 2005. Digitalna knjižnica Slovenije – dLib.si. [The digital library of Slovenia – dLib.si] *Informatika kot temelj povezovanja: zbornik posvetovanja*, pp. 683-689.

Martina Orožen. 1996. *Oblikovanje enotnega slovenskega knjižnega jezika v 19. stoletju.* [The formation of a unified Slovene literary language in the XIXth Century.] Ljubljana, Filozofska fakulteta.

Erich Prunč. 2007. Deutsch-slowenische/kroatische Übersetzung 1848-1918. Ein Werkstättenbericht. [German-Slovene/Croatian translation, 1848-1918. Workshop report]. *Wiener Slavistisches Jahrbuch 53/2007*. Austrian Academy of Sciences Press, Vienna. pp. 163-176.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicolas Smith, 2007. Tagging the Bard: Evaluating the accuracy of a modern PoS tagger on Early Modern English corpora. In Proceedings of Corpus Linguistics 2007. Uni. of Birmingham, UK.

Ulrich Reffle, Efficiently generating correction suggestions for garbled tokens of historical language, Journal of Natural Language Engineering, Special Issue on Finite State Methods and Models in Natural Language Processing, 2011.

T2HSOM: Understanding the Lexicon by Simulating Memory Processes for Serial Order

Marcello Ferro, Claudia Marzi, Vito Pirrelli
Institute for Computational Linguistics "A. Zampolli", National Research Council
via G. Moruzzi 1, Pisa, Italy
e-mail: {marcello.ferro, claudia.marzi, vito.pirrelli}@ilc.cnr.it

Abstract

Over the last several years, both theoretical and empirical approaches to lexical knowledge and encoding have prompted a radical reappraisal of the traditional dichotomy between lexicon and grammar. The lexicon is not simply a large waste basket of exceptions and sub-regularities, but a dynamic, possibly redundant repository of linguistic knowledge whose principles of relational organization are the driving force of productive generalizations. In this paper, we overview a few models of dynamic lexical organization based on neural network architectures that are purported to meet this challenging view. In particular, we illustrate a novel family of Kohonen self-organizing maps (T2HSOMs) that have the potential of simulating competitive storage of symbolic time series while exhibiting interesting properties of morphological organization and generalization. The model, tested on training samples of as morphologically diverse languages as Italian, German and Arabic, shows sensitivity to manifold types of morphological structure and can be used to bootstrap morphological knowledge in an unsupervised way.

1. Introduction

Traditional generative approaches to language inquiry view word competence as consisting of a morphological lexicon, an assorted hotchpotch of exceptions and sub-regularities, and a grammar, a set of productive combinatorial rules (Di Sciullo and Williams 1987; Prasada and Pinker 1993). Whatever cannot be assembled through rules must be relegated wholesale to the lexicon, whose size depends on the generative power of the grammar: the richer the power, the poorer the lexicon.

Baayen (2007) observes that the approach reflects an outdated view of lexical storage as more 'costly' than computational operations. Similarly, alternative theoretical models question the primacy of grammar rules over lexical storage, arguing that morphological regularities emerge from independent principles of lexical organization, whereby fully inflected forms are redundantly stored and mutually related through entailment lexical relations (Matthews 1991; Pirrelli 2000; Burzio 2004; Blevins 2006). This view prompts a radically different computational metaphor than traditional generative models. A speaker's knowledge corresponds more to one large dynamic relational database than to a general-purpose automaton augmented with lexical storage.

In spite of the large body of theoretical literature on the topic, however, few computational models of the lexicon can be said to address such a complex interaction between storage and computation. Contrary to what is commonly held, connectionism has failed to offer an alternative view of the interplay between lexicon and grammar. As we shall argue in more detail in the ensuing session, there is no place for the lexicon in classical connectionist networks. Somewhat ironically, they seem to have adhered to a cornerstone of the rule-based approach to morphological inflection, thus providing a neurally-inspired mirror image of inflection rules.

In this paper, we will explore the somewhat complementary view that storage plays a fundamental role in lexical modelling, and that computer simulations of short-term and long-term memory processes can go a long way in addressing issues of lexical organization. The present paper lends support to this claim by illustrating a novel neural network architecture known as "Topological Temporal Hebbian Self-Organizing Map" (or T2HSOM for short, Ferro *et al.* 2010). A T2HSOM has the potential of simulating dynamic storage of symbolic time series while exhibiting interesting properties of morphological self-organization. Trained on morphologically diverse families of word forms, T2HSOMs can be shown to bootstrap morphological structure in an unsupervised way. Finally, we suggest that they offer an ideal workbench for understanding the structure of the lexicon by simulating memory processes.

2. Background

As a first approximation, the lexicon is the store of words in long-term memory. Any attempt at modelling lexical competence must hence take issues of string storage very seriously. In this respect, the rich cognitive literature on short-term and long-term memory processes (Miller 1956; Baddeley and Hitch 1974; Baddeley 1986; 2006; Henson 1998; Cowan 2001; among others) has the unquestionable merit of highlighting some fundamental issues of coding, maintenance and manipulation of time-bound constraints over strings of symbols.

Word forms are primarily sequences of sounds or letters and so the question of their coding (and maintenance) in time is logically prior to any other processing issue. In spite of this truism, however, coding issues have suffered unjustified neglect by the NLP research community over the last 30 years. In fact, the mainstream connectionist answer to the problem of time series coding, namely so-called "conjunctive coding", appears to elude some core issues in lexical representation.

Conjunctive codes (e.g., Coltheart, Rastle, Perry, Langdon and Ziegler 2001; Harm and Seidenberg 1999; McClelland and Rumelhart 1981; Perry, Ziegler, and

Zorzi 2007; Plaut, McClelland, Seidenberg, and Patterson 1996) are typically assumed to be available in the input (or encoding) layer of a multi-layered perceptron in the form of a built-in repertoire of context-sensitive Wickelphones, such as $_{\#}C_{a}$ and $_{c}A_{t}$ to respectively encode the letters *c* and *a* in *cat*. However, the use of Wickelphones raises the immediate issue of their ontogenesis, since they appear to solve the problem of coding time series by resorting to time-bound relations whose representation in the encoding layer remain unexplained. A second related issue is the acquisition of phonotactic knowledge. Speakers are known to exhibit differential sensitivity to diverse sound patterns. Effects of graded specialization in the discrimination of sound clusters and lexical well-formedness judgements are the typical outcome of acquiring a particular language. If such patterns are part and parcel of the encoding layer, the same processing system cannot be used to deal with different languages exhibiting differential sound constraints.

A third limitation of conjunctive coding is that phonemes and letters are bound with their context. This means that two elements like $_{\#}E_{v}$ and $_{v}E_{r}$ representing two instances of the same letter *e* in *#every* are in fact as similar (or as different) as any two other elements. We are just left with token representations, the notion of type of unit remaining out of the representational reach of the system. This makes it difficult to generalize knowledge about phonemes or letters across positions (the so-called dispersion problem: Plaut, McClelland, Seidenberg, and Patterson 1996; Whitney 2001). It is also difficult to align positions across word forms of differing lengths (i.e., the alignment problem: see Davis and Bowers 2004), thus hindering recognition of both shared and different sequences between morphologically-related forms. The failure to provide a principled solution to alignment problems (Daugherty and Seidenberg 1992; Plaut, McClelland, Seidenberg, and Patterson 1996; Seidenberg and McClelland 1989) is particularly critical from the perspective of lexical storage. Languages wildly differ in the way morphological information is sequentially encoded, ranging from suffixation to prefixation, sinaffixation, apophony, reduplication, interdigitation and combinations thereof. For example, the alignment of lexical roots in three as diverse pairs of paradigmatically related forms such as English ***walk-walk****ed*, Arabic ***k****a****t****a****b****a****-y****a****k****t****u****b****u*** (‘he wrote’ - ‘he writes’), German ***mach****en-ge****mach****t* (‘make’-‘made’ past participle) requires substantially different processing strategies. Pre-coding any such strategy into lexical representations (e.g. through a fixed templatic structure that separates the lexical root from other morphological markers) would have the effect of slipping in morphological structure directly into the input, thereby making input representations dependent on languages. A far more plausible solution would be to let the processing system home in on the right sort of alignment strategy through repeated exposure to a range of language-specific families of morphologically-related words. This is exactly what conjunctive coding cannot do.

To our knowledge, there have been three attempts to tackle the issue within a connectionist framework: Recursive Auto-Associative Memories (RAAM; Pollack 1990), Simple Recurrent Networks (SRN; Botvinick and Plaut 2006) and Sequence Encoders (Sibley et al. 2008). The three models set themselves different goals: i) encoding an explicitly assigned hierarchical structure for RAAM, ii) simulation of a range of behavioural facts of human Immediate Serial Recall for Botvinick and Plaut’s SRNs and iii) long-term lexical entrenchment for the Sequence Encoder of Sibley and colleagues.

In spite of their considerable differences, all systems share the important feature of modelling storage of symbolic sequences as the by-product of an auto-encoding task, whereby an input sequence of arbitrary length is eventually reproduced on the output layer after being internally encoded through recursive distributed patterns of node activation on the hidden layer(s). Serial representations and memory processes are thus modelled as being contingent on the task. In particular, Botvinick and Plaut’s paper makes the somewhat paradoxical suggestion that human performance on immediate serial recall develops through direct practice on the task of word repetition. Moreover, short-term memory effects appear to be accounted for in terms of a long-term dynamics dictated by the process of weight adjustment through learning. Although long-term memory effects are known to increase short-term storage capacities, developmental evidence shows that the causal relationship is in fact reversed, with children with higher order short-term memory being able to hold on to new words for longer, thus increasing the likelihood of long-term lexical learning (Baddeley 2007).

We describe here a novel computational architecture for lexical processing and storage. The architecture is based on Kohonen’s Self-Organizing Maps (SOMs; Kohonen 2001) augmented with first-order associative connections that encode probabilistic expectations (so called, Topological Temporal Hebbian SOMs, or T2HSOMs for short; Koutnik 2007; Pirrelli et al. in press; Ferro et al. 2010). T2HSOMs mimic the behaviour of brain maps, medium to small aggregations of neurons in the cortical area of the brain, involved in selectively processing homogeneous classes of data. T2HSOMs define an interesting class of general-purpose memory models for serial order, exhibiting a non-trivial interplay between short-term and long-term memory processes. At the same time, they simulate incremental processes of topological self-organization whereby lexical sequences are arranged in maximally predictive hierarchies exhibiting interesting morphological structures.

3. Topological Temporal SOMs

T2HSOMs are grids of topologically organized memory nodes with dedicated sensitivity to time-bound stimuli. Upon presentation of an input stimulus, all map nodes are activated synchronously, but only the most highly activated one, the so-called Best Matching Unit (BMU), wins over the others. Figure 1 illustrates two chains of BMUs triggered by the input German forms *gemacht* and *gelacht* (‘made’ and ‘laughed’ past participle) exposed to a 20x20

nodes map one letter at a time. In the Figure, each node is labelled with the letter the node is most sensitive to after training. Pointed arrows represent temporal connections linking two consecutively activated nodes. The thickness of each arrow gives the strength of the temporal connection. Finally, arrows depict the temporal sequence of node exposure (and node activation), starting from the beginning-of-the-word symbol '#' (anchored in the top left corner of the map) and ending with '$'.

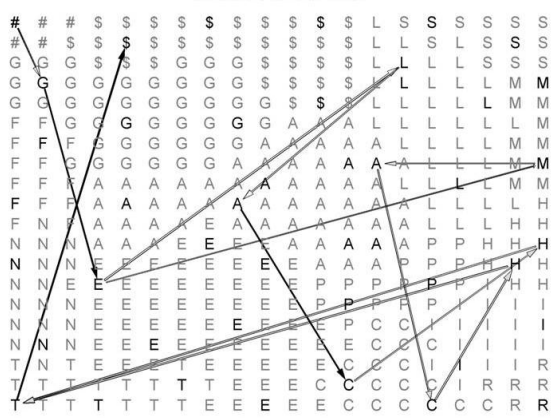


Figure 1 – BMU activation chains for *gemacht-gelacht*

Dedicated sensitivity and topological organization are not wired-in on the map. Neighbouring nodes become increasingly sensitive to letters that are similar in both encoding and distribution through drilling.

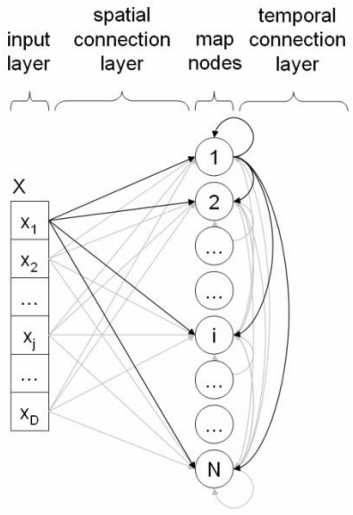


Figure 2 - Outline architecture of a T2HSOM

Figure 2 offers the architecture of a T2HSOM. Each node in the map is connected with all elements of the input layer through communication channels with no time delay, whose strength is modified through training. Connections on the temporal layer, on the other hand, are updated with a fixed one-step time delay, based on activity synchronization of the BMU at time $t-1$ and the BMU at time t. It is important to appreciate at this juncture that, unlike classical conjunctive representations in either Simple Recurrent Networks (Elman 1991) or Recursive SOMs (Voegtlin 2002), where both order and item information is collapsed on the same layer of connectivity, T2HSOMs keep the two sources of information stored on separate (spatial and temporal) layers, which are trained according to independent principles. The aspect has interesting repercussions on issues of order-independent generalizations over symbol types and goes a long way to addressing both dispersion and alignment problems in word matching.

3.1 Memory structures and memory orders

Through repeated exposure to word forms encoded as time series of letters, a T2HSOM shows a tendency to dynamically store strings as trie-like graphs, eliminating prefix redundancy and branching out when two (or more) different nodes are alternative continuations of the same history of past activated nodes (Figure 1). This lexical organization accords well with cohort models of lexical access (Marslen Wilson 1987) and is in keeping with a wide range of empirical evidence on human word processing and storage: i) development of minimally-entropic forward chains of linguistic units, enhancing predictive and anticipatory behaviour in language processing (Altmann and Kamide 1999; Federmeier 2007; Pickering and Garrod 2007); ii) frequency-based competition between inflected forms of the same lexical base (e.g. *brings* and *bringing*) (Hay 2001; Ford, Marslen-Wilson and Davis 2003; Lüdeling and De Jong 2002; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder and Baayen 2004); iii) simultaneous activation of false morphological friends (e.g. *broth* and *brother*) (Frost et al. 1997; Longtin et al. 2003; Rastle et al. 2004; Post, Marslen-Wilson, Randall and Tyler 2008).

It can be shown that trie-like memory structures maximize the map's expectation of upcoming symbols or, equivalently, minimize the entropy over the set of transition probabilities between consecutive BMUs. This is achieved through a profligate use of memory resources, whereby several nodes are recruited to be most sensitive to contextually specific occurrences of the same letter.

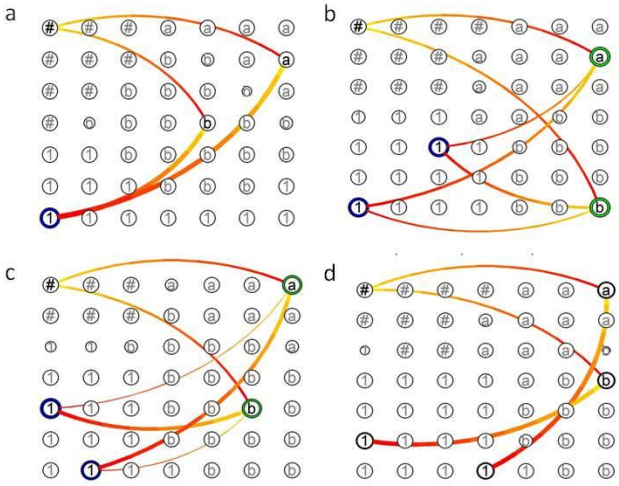


Figure 3 – Stages of chain dedication through learning

Figure 3 illustrates how this process of incremental specialization unfolds through training. For simplicity we are assuming that the map is trained on two strings only: *#a1* and *#b1*. Panel a) represents an early stage of learning, when the map recruits a single BMU for the symbol *1* irrespective of its embedding context. After some more learning epochs, two BMUs are recruited after an *a* or a *b* through equally strong connections (Panel b). Connections get increasingly specialized in Panel c), where the two *1* nodes are preferentially selected by either context. Finally, Panel d) illustrates a stage of dedicated connections, where each *1* node is selected by one specific left context only. This stage is reached when the map can train each single node without affecting any neighbouring node. Technically, this corresponds to a learning stage where the map's neighbourhood radius *r* is equal to 0.

4. Emergent Morphological Structure

To what extent do we find morphological structure in a lexical map organized according to the principles sketched above? We observe a straightforward correlation between morphological segmentation and topological organization of BMUs on the map: word forms sharing sub-lexical constituents tend to trigger chains of identical or neighbouring nodes.

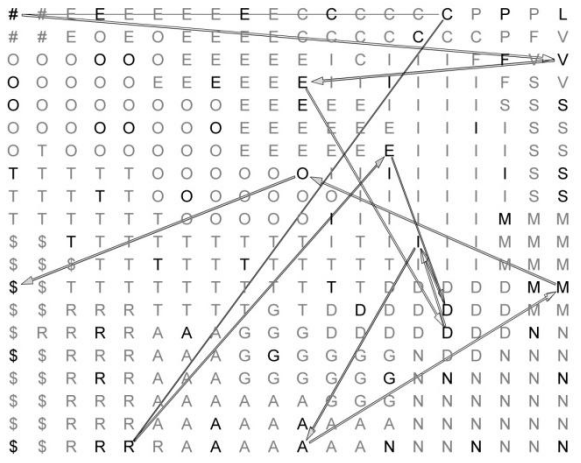


Figure 4 – BMU activation chains for *crediamo-vediamo*

The map distance between BMUs triggered by identical morphemic constituents of two morphologically-related forms is expected to be shorter than the map distance between BMUs activated by morphologically heterogeneous constituents. In a nutshell, topological distance is a function of morphological proximity. In traditional approaches to word segmentation, this is equivalent to aligning morphologically-related word forms by morphological structure. As chains of activated nodes encode time sequences of symbols, T2HSOMs can be said to enforce alignment through synchrony.

To illustrate, we trained three different instances of a T2HSOM on Italian, German and Arabic verb forms. Figure 4 plots the activation chains of the present indicative forms *vediamo* ('we see') and *crediamo* ('we believe') on a 20x20 nodes Italian map, trained on 32 Italian verb forms. The chains are clearly separated on the roots *cred-* and *ved-*, but converge as soon as more letters are shared by the two forms. Eventually the substring *-iamo* activates a unique BMU chain. We take this to mean that the substring is recognized by the map as encoding the same type of inflectional ending. Note that the shared substring *-iamo* takes different positions in the two forms, starting from the forth letter in *vediamo* and from the fifth letter in *crediamo*. In traditional positional coding, this raises an alignment problem. In our map, *-iamo* receives a converging topological representation, as order information is relative and time-dependent rather than absolute.

German past participles provide a case of discontinuous morphological structure. Let us turn back to Figure 1 above. Note that *gemacht* and *gelacht* share the same sequence of BMUs for *ge-*, but they part on the roots *mach-* and *lach-* to eventually meet again upon recognition of the ending *–t*. This is expressed in terms of topological distance between BMUs in Figure 5, giving the per-node topological distance of the BMU chains for *gemacht* and *gelacht*.

	#	G	E	L	A	C	H	T	$
#	0.00	0.24	0.43	0.44	0.40	0.58	0.59	0.53	0.25
G	0.24	0.00	0.37	0.42	0.35	0.52	0.54	0.48	0.24
E	0.43	0.37	0.00	0.47	0.30	0.37	0.48	0.30	0.40
M	0.55	0.52	0.49	0.31	0.38	0.42	0.25	0.59	0.48
A	0.45	0.41	0.39	0.27	0.10	0.38	0.31	0.51	0.38
C	0.63	0.57	0.43	0.49	0.40	0.06	0.34	0.46	0.57
H	0.58	0.53	0.46	0.38	0.37	0.33	0.03	0.54	0.51
T	0.53	0.48	0.30	0.59	0.42	0.40	0.56	0.00	0.52
$	0.25	0.24	0.40	0.36	0.34	0.53	0.51	0.52	0.00

Figure 5 – Topological distance matrix for *gemacht-gelacht*

Besides identical nodes for *ge-* and *–t*, the matrix shows that morphological structure is inherently graded on morpheme boundaries, with the topological distance between the roots narrowing down as the shared suffix gets closer, in keeping with psycholinguistic evidence on word processing (Hay and Baayen 2005).

	#	G	E	S	P	I	E	L	T	$
#	0.00	0.24	0.43	0.46	0.55	0.63	0.57	0.44	0.53	0.25
S	0.51	0.50	0.53	0.06	0.41	0.44	0.55	0.26	0.61	0.44
P	0.52	0.47	0.35	0.44	0.06	0.31	0.29	0.38	0.37	0.46
I	0.63	0.58	0.48	0.47	0.26	0.00	0.38	0.42	0.48	0.56
E	0.57	0.51	0.16	0.55	0.33	0.38	0.00	0.49	0.29	0.52
L	0.44	0.42	0.45	0.24	0.36	0.42	0.49	0.00	0.54	0.37
E	0.42	0.36	0.08	0.43	0.33	0.41	0.15	0.37	0.34	0.37
N	0.48	0.42	0.24	0.57	0.44	0.51	0.34	0.52	0.24	0.46
$	0.39	0.36	0.40	0.28	0.36	0.43	0.46	0.24	0.49	0.14

Figure 6 – Topological distance matrix for *spielen-gespielt*

A case of root-alignment in German lexically-related forms is illustrated in Figure 6, showing the per-node distance between *spielen* and *gespielt*. Once more, this would be out of reach of positional coding.

More difficult cases of root-alignment arise in the context of Semitic morphologies, where the relative position of the letters shared by lexically-related forms vary dramatically, as in *kataba* vs. *yaktubu*, respectively the perfective and imperfective forms of the verb triliteral root *ktb* ('write'). An interesting related question is to what extent the activation chains corresponding to Arabic perfective and imperfective forms are successful in representing the morphological notions of triconsonantal root and interdigitated vowel pattern. The problem is not trivial, as discontinuous morphological patterns are known to be beyond the reach of chaining models for serial order. Given two forms like *kataba* ('he wrote') and *hadama* ('he shattered') for example, vowels in the two strings are all preceded by different left contexts.

	#	K	a	T	a	B	a
#	0.00	0.41	0.28	0.63	0.55	0.46	0.49
H	0.44	0.22	0.37	0.32	0.30	0.26	0.39
a	0.30	0.34	0.05	0.53	0.25	0.40	0.20
D	0.57	0.31	0.51	0.19	0.28	0.30	0.47
a	0.60	0.43	0.34	0.32	0.05	0.44	0.24
M	0.49	0.21	0.47	0.30	0.39	0.17	0.53
a	0.53	0.48	0.25	0.51	0.18	0.53	0.05

Figure 7 – Topological distance matrix for *kataba-hadama*

Figure 7 illustrates the solution offered by a T2HSOM to the problem. The three *a*'s in the perfective vowel pattern are given dedicated representations on the map, triggering differently located BMUs. Not only is the map able to discriminate between three different instances of the same symbol (*a*) in the same string (*kataba*), but it can also align each such *a* with its homologous *a* in another morphologically-related form (*hadama*). In fact, this seems to be a necessary step to take if we want the map to get a notion of the Arabic perfective vowel pattern.

To understand how this is possible, observe that temporal information is not limited to information about the actually occurring left context. The BMU activated by the symbol *a* in the input string *#ha* at time *t* receives support, through temporal connections, from all nodes activated at time *t-1*. The nodes include, among others, the *k* node, which competes with the *h* node at time *t-1* as it receives temporal support from the # node activated at time *t-2* (due to the existence of *#ka* in *kataba*). By reverberating simultaneous activation of competing nodes to an ensuing state, the map can place *a* nodes triggered by *#ka* and *#ha* in the same area, as they share a comparatively large portion of pre-synaptic support. In general, the mechanism allows the map to keep together nodes activated by letters in the same position in the string.

5. Lexical access and recall

So far, we considered chains of BMU activation based on exposure to time-bound sequences of letters. By inspecting activation chains, we can tell whether the map recognizes an input signal as a specific sequence of symbols or not. This is not trivial and requires both sensitivity to letter codes and the capacity of anticipating upcoming symbols on the basis of already seen symbols. Nonetheless, it says little about issues of lexical storage *per se*. How do we know that the map has actually stored the sequence it is able to recognize?

We can model lexical recall as the task of reinstating a sequence of letters from the integrated pattern of activation of a map that has just seen that sequence. Recall that a form is exposed to the map one letter at a time. At each time tick, each letter leaves an activation pattern that accumulates in the map short-term buffer. When the whole form is shown, the map's short-term buffer will thus retain the concurrent activation of all letters forming the just seen word (Figure 8).

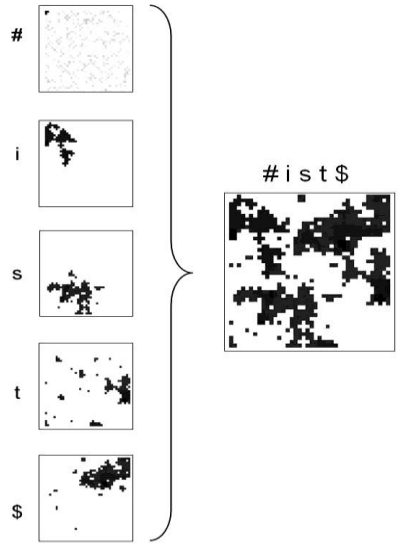


Figure 8 – Per-letter and concurrent activation for *#ist$*

We may eventually feed this pattern back into the map and ask the map to recall from it the expected sequence of letters. Note that this is a considerably more difficult task than activating a specific node upon seeing a particular letter. A whole word integrated pattern of activation is the lexical representation for that word. If the map is able to accurately encode letters and their order of appearance, it will be successful in accessing and retrieving the whole word from its long-term store.

To assess the capacity of a T2HSOM to develop, access and retrieve lexical representations, we trained a 40x40 map on 5000 Italian word forms, sampled from the book *The Adventures of Pinocchio* by Collodi. We then probed the memory content of the map on two test sets: the entire set of "training" word tokens (about 1050 different form types), and a sample of about 250 unseen inflected forms of all verbs that are found in the training set in at least one other form. No frequency information was given for the latter "testing" set.

Results of the experiments are shown in Figure 9 in terms of per-word and per-letter accuracy over types and tokens.

Italian			accuracy	
			% types	% tokens
recognition	training set	per word	99.2	99.7
		per letter	100	100
	testing set	per word	99.6	99.6
		per letter	100	100
recall	training set	per word	97.3	98.8
		per letter	99.1	99.6
	testing set	per word	75.7	75.7
		per letter	95.1	95.1

Figure 9 – Accuracy results on seen and unseen Italian word forms

German			accuracy	
			% types	% tokens
recognition	training set	per word	99.6	98.5
		per letter	99.6	99.9
	testing set	per word	96.7	96.7
		per letter	99.6	99.6
recall	training set	per word	94.2	97.9
		per letter	98.9	99.6
	testing set	per word	80.7	80.7
		per letter	95.8	95.8

Figure 10 – Accuracy results on seen and unseen German word forms

Figure 10 shows the results of a 40x40 T2HSOM trained on 5000 German word tokens (about 1750 different form types), sampled from three fairy tales by brothers Grimm. The testing set included 150 unseen inflected forms of verbs and nouns that are found in the training set in at least one other form, with no frequency information.

All in all, T2HSOMs show a remarkable capacity of activating appropriate BMUs upon recognition of input letters, both on seen words (training set) and unseen words (testing set). Moreover, they can also recall most such words. In fact more than 97% of the Italian forms and more than the 94% of the German forms in the training set are retrieved accurately through activation of BMUs chains. On both the Italian and German training sets, recall errors strongly correlate with low word frequency and word length effects, with most missed word forms showing frequency values close to 1 (Figure 11).

That more than just storage is involved here is shown by the results on the testing set, assessing the ability of the map to "recall" unseen words. More than 75% Italian unseen words and 80% German unseen words are retrieved accurately, meaning that the maps developed memory traces of expected, rather than simply attested, sequences. T2HSOMs can in fact structure familiar information in a very compact (but accurate) way through shared activation paths, thus making provision for connection chains that are never triggered in the course of training. The effect is reminiscent of what we noted in Figure 3 above, where wider neighbourhoods, typical of early stages of learning, favour profligate and more liberal inter-node connections. Only when the map is free to train neighbouring nodes independently, dedicated paths develop. In the current experimental setting, the map is too small to be able to dedicate a different node to each different context-dependent occurrence of a letter.[1] Fewer nodes are recruited to be sensitive to several different context-dependent tokens of the same letter type and to be more densely connected with other nodes. A direct consequence of this situation is generalization, corresponding to the configurations shown in 3.b) and 3.c), where both the *a* and *b* nodes develop more outgoing connections than those strictly required by the training evidence. Most notably, this is the by-product of the way the map stores and structures lexical information.

Italian training set	frequency		length	
	μ	σ	μ	σ
all words	2.8	7.4	7.0	2.5
correctly recalled words (97.3%)	2.8	7.5	7.0	2.5
wrongly recalled words (2.7%)	1.2	0.4	8.6	2.4
German training set				
all words	2.9	6.7	5.9	2.4
correctly recalled words (94.2%)	3.0	6.9	5.7	2.3
wrongly recalled words (5.8%)	1.1	0.3	8.9	2.7

Figure 11 – Mean value and standard deviation of word form frequency and length for Italian and German training sets.

6. Concluding Remarks and Developments

To date, both symbolic and connectionist approaches to the lexicon have laid emphasis on processing aspects of word competence only, whereby morphological productivity is modelled as the task of outputting a – possibly – unknown word form (say an inflected form like *shook*) by taking as input its lexical base (*shake*). Such a "derivational" approach to word competence (Baayen 2007), however, obscures the interplay between storage and computation, adhering to a view of morphological competence as the ability to play a word game.

Symbolic approaches encode word forms using traditional computational devices for storage, allocation and serial order representation such as ordered sets, strings and the like. These devices provide built-in means of serializing order information through chains of pointers which are accessed and manipulated by independently required recursive algorithms. In classical connectionist architectures (Rumelhart and McClelland 1986), on the other hand, the internal representation of word forms in the lexicon is modelled by the pattern of connections between the hidden and the output layer in a multilayered

[1] A 1600 nodes T2HSOM uses up the 2.5% level of connectivity required to store all forms as dedicated BMU chains.

perceptron mapping lexical bases onto inflected forms (e.g. *go* vs. *went*). Serial order is pre-encoded through dedicated nodes, and the resulting lexical organization appears to be contingent upon the requirements of the task of generating novel forms. In principle, different tasks may impose different structures on the lexicon.

In this paper we took a somewhat different approach to the problem. We assumed that word storage plays a fundamental role in both word learning and processing. The way words are structured in our long-term memory (the lexicon) is key to understanding the mechanisms governing word processing and productivity. This perspective offers a few advantages. First, it allows scholars to properly focus on word productivity (the *explanandum*) as the by-product of more basic memory strategies (our *explanans*) that must independently be assumed to account for fundamental aspects of word learning (including but not limited to storage of word forms). Secondly, it opens up new promising avenues of scientific inquiry by tapping the large body of empirical evidence on short-term and long-term memorization strategies for serial order (see Baddley 2007 for a comprehensive recent overview). Furthermore, it gives the opportunity of using sophisticated computational models of language-independent memory processes (Brown Preece and Hulme 2000; Henson 1998; Burgess and Hitch 1996, among others) to shed light on language-specific aspects of word encoding. Finally, it promises to provide a comprehensive picture of the complex dynamics between computation and memory underlying morphological processing.

Put in a nutshell, the processing of unknown words requires mastering rule-governed combinatorial processes. In turn, these processes presuppose knowledge of the sub-word units to be combined. We argue that preliminary identification of the basic inventory of such units depends on memorization of their complex combinations. The way information is stored thus reflects the way such information is dynamically represented, and eventually accessed and retrieved as patterns of concurrent activation of memory areas. According to the view endorsed here, memory processes have the ability not only to hold information but also to structure and manipulate it.

By exploiting the full potential of T2HSOMs, we can simulate processes of dynamic interaction between short-term and long-term memory processes on a classical memory task like Immediate Serial Recall (Henson 1998; Cowan 2001). Moreover, we can investigate aspects of co-organization of concurrent temporal maps, each trained on different modalities of the same input stimuli. This dynamic is key to modelling pervasive aspects of synchronization of multi-modal sequences in both linguistic (e.g. reading) and extra-linguistic (e.g. visuomotor coordination) tasks (Ferro et al. 2011). Finally, we are in a position to explore emergence of islands of reliability (Albright 2002) in the morphological lexicon to account for processes of analogy-driven generalization on the morphological input.

7. References

Albright, Adam (2002). 'Islands of reliability for regular morphology: Evidence from Italian', *Language* 78: 684-709.

Altmann, G.T.M., and Kamide, Y. (1999), Incremental interpretation at verbs: restricting the domain of subsequent reference, *Cognition*, 73, 247-264

Baayen, H. (2007), Storage and computation in the mental lexicon, in G. Jarema and G. Libben (eds.), *The Mental Lexicon: Core Perspectives*, Amsterdam, Elsevier, 81-104.

Baddeley, A.D. (1986), *Working memory*, New York, Oxford University Press.

Baddeley, A.D. (2006), Working memory: an overview, in S. Pickering (ed.), *Working Memory and Education*, New York, Academic Press, 1-31.

Baddeley, A.D. (2007), *Working memory, thought and action*, Oxford, Oxford University Press.

Baddeley, A.D., and Hitch, G. (1974), Working memory, in G.H. Bower (ed.), *The psychology of learning and motivation: Advances in research and theory*, New York, Academic Press, 8, 47-89.

Blevins, J.P. (2006), Word-based morphology, *Journal of Linguistics*, 42, 531-573.

Botvinick, M., and Plaut, D.C. (2006), Short-term memory for serial order: A recurrent neural network model, *Psychological Review*, 113, 201-233.

Burzio, L. (2004), Paradigmatic and syntagmatic relations in italian verbal inflection, in J. Auger, J.C. Clements and B. Vance (eds.), *Contemporary Approaches to Romance Linguistics*, Amsterdam, John Benjamins.

Cowan, N. (2001), The magical number 4 in short-term memory: A reconsideration of mental storage capacity, *Behavioral and Brain Sciences*, 24, 87-185.

Daugherty, K., and Seidenberg, M.S. (1992), Rules or connections? The past tense revisited, in *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ, Erlbaum.

Davis, C.J., and Bowers, J.S. (2004), What do Letter Migration Errors Reveal About Letter Position Coding in Visual Word Recognition?, *Journal of Experimental Psychology: Human Perception and Performance*, 30, 923-941.

Di Sciullo, A. M. and Williams, E. (1987). *On the Definition of Word*, Cambridge, MA: MIT Press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001), DRC: A Dual Route Cascaded model of visual word recognition and reading aloud, *Psychological Review*, 108, 204-256.

Elman, J.L. (1990), Finding Structure in Time, *Cognitive Science*, 14(2), 179-211.

Federmeier, K.D. (2007), Thinking ahead: the role and roots of prediction in language comprehension, *Psychophysiology*, 44, 491-505.

Ferro, M., Ognibene, D., Pezzulo, G., and Pirrelli, V. (2010), Reading as active sensing: a computational model of gaze planning in word recognition, *Frontiers in Neurorobotics*, DOI: 10.3389/fnbot.2010.00006,

issn: 1662-5218, 4(6), 1-16.

Ferro, M., Chersi, F., Pezzulo, G., and Pirrelli, V. (2011), Time, Language and Action - A Unified Long-Term Memory Model for Sensory-Motor Chains and Word Schemata, in *Intelligent and Cognitive systems*, P. Kunz (ed.), *ERCIM News*, vol. 84 pp. 27-28.

Ford, M., Marslen-Wilson, W., and Davis, M. (2003), Morphology and frequency: contrasting methodologies, in H. Baayen and R. Schreuder (eds.), *Morphological Structure in Language Processing*, Berlin-New York, Mouton de Gruyter.

Frost, R., Forster, K.I., and Deutsch, A. (1997), What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, 829-856.

Harm, M.W., and Seidenberg, M.S. (1999), Phonology, Reading Acquisition and Dyslexia: Insights from Connectionist Models, *Psychological Review*, 106(3), 491-528.

Hay, J. (2001), Lexical frequency in morphology: is everything relative?, *Linguistics*, 39, 1041-1111.

Hay, J.B., and Baayen, R.H. (2005), Shifting paradigms: gradient structure in morphology, *Trends in Cognitive Sciences*, 9, 342-348.

Henson, R.N. (1998), Short-term memory for serial order: The start-end model, *Cognitive Psychology*, 36, 73-137.

Kohonen, T. (2001), *Self-Organizing Maps*, Heidelberg, Springer-Verlag.

Koutnik, J. (2007), Inductive Modelling of Temporal Sequences by Means of Self-organization, in *Proceeding of International Workshop on Inductive Modelling* (IWIM 2007), Prague, 269-277.

Lüdeling, A., and Jong, N. de (2002), German particle verbs and word formation, in N. Dehé , R. Jackendoff, A. McIntyre and S. Urban, (eds.), *Explorations in Verb-Particle Constructions*, Berlin, Mouton der Gruyter.

Marslen-Wilson, W. (1987), Functional parallelism in spoken word recognition, *Cognition*, 25, 71-102.

Matthews, P.H. (1991), *Morphology*, Cambridge, Cambridge University Press.

McClelland, J.L., and Rumelhart, D.E. (1981), An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings, *Psychological Review*, 88, 375-407.

Miller, G.A. (1956), The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63 (2), 81-97.

Moscoso del Prado Fermìn, M., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, H. (2004), Morphological Family Size in a Morphologically Rich Language: The Case of Finnish Compared With Dutch and Hebrew, «Journal of Experimental Psychology: Learning, Memory and Cognition», 30(6), 1271-1278.

Perry, C., Ziegler, J. C., and Zorzi, M. (2007), Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud, *Psychological Review*, 114(2), 273-315.

Pickering, M.J. and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11, 105-110

Pirrelli, V. (2000), *Paradigmi in Morfologia. Un approccio interdisciplinare alla flessione verbale dellitaliano*, Pisa, Istituti Editoriali e Poligrafici Internazionali.

Pirrelli, V., Ferro, M., and Calderone, B. (in press), Learning paradigms in time and space. Computational evidence from Romance languages, in M. Goldbach, M.O. Hinzelin, M. Maiden and J.C. Smith (eds.) *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, Oxford, Oxford University Press.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., and Patterson, K. (1996), Understanding normal and impaired word reading: Computational principles in quasi-regular domains, *Psychological Review*, 103, 56-115.

Pollack, J. B. (1990), Recursive distributed representations, *Artificial* Intelligence, 46, 77–105.

Post, B., Marslen-Wilson, W., Randall, B., and Tyler, L.K. (2008), The processing of English regular inflections: Phonological cues to morphological structure, *Cognition*, 109, 1-17.

Prasada, S., and Pinker, S. (1993), Generalization of regular and irregular morphological patterns, *Language and Cognitive Processes* 8, 1-56.

Rastle, K., Davis and M.H. (2004), The broth in my brothers brothel: Morpho-orthographic segmentation in visual word recognition, *Psychonomic Bulletin and Review*, 11(6), 1090-1098.

Seidenberg, M.S., and McClelland, J.L. (1989), A distributed, developmental model of word recognition and naming, in A. Galaburda (ed.), *From neurons to reading*, MIT Press.

Sibley, D.E., Kello, C.T., Plaut, D., and Elman, J.L. (2008), Large-scale modeling of wordform learning and representation, *Cognitive Science*, 32, 741–754.

Voegtlin, T. (2001) Recursive self-organizing maps, *Neural Networks*, 15, 979–991

Whitney, C. (2001), How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review, *Psychonomic Bulletin and Review*, 8, 221-243.

Appendix - The *T2HSOM* model

A.1 Short-term dynamics: activation and filtering

In recognition mode, the activation level of the map's *i*-th node at time *t* is:

$$y_i(t) = \alpha \cdot y_{S,i}(t) + \beta \cdot y_{T,i}(t)$$

where α and β weigh up the respective contribution of the spatial and temporal layers, and

$$y_{S,i}(t) = \sqrt{D} - \sqrt{\sum_{j=1}^{D} [x_j(t) - w_{i,j}(t)]^2}$$

is the normalized Euclidean distance between the input vector *x(t)* at time *t* and the spatial weight vector associated with the *i-th* node, and

$$y_{T,i}(t) = \sum_{h=1}^{N} [y_h(t-1) \cdot m_{i,h}(t)]$$

is the weighted temporal pre-activation of the *i*-th node at time *t* prompted by the state of activation of all *N* nodes of the map at time *t-1*. The BMU at time *t* is identified by looking for the maximum activation level

$$y'_{bmu}(t) = \max_i \{y'_i(t)\}$$

eventually normalized to ensure network stability over time:

$$Y(t) = \frac{Y'(t)}{y'_{bmu}(t)}$$

A.2 Long-term dynamics: learning

In *T2HSOM* learning consists in topological and temporal co-organization.

(i) Topological learning

In classical SOMs, this effect is taken into account by a neighbourhood function centered around BMU. Nodes that lie close to BMU on the map are strengthened as a function of BMU's neighbourhood. The distance between BMU and the *i*-th node on the map is calculated through the following Euclidean metrics:

$$d_i(t) = \sqrt{\sum_{c=1}^{n} [i_c - bmu_c(t)]^2}$$

where *n* is 2 when the map is two-dimensional. The topological neighbourhood function of the *i*-th neuron is defined as a Gaussian function with a cut-off threshold:

$$c_{S,i}(t) = \begin{cases} e^{-\frac{d_i^2(t)}{2\sigma_S^2(t_E)}} & if\ d_i(t) \le \nu_S(t_E) \\ 0 & if\ d_i(t) > \nu_S(t_E) \end{cases}$$

where $\sigma_S(t_E)$ is the topological neighbourhood shape coefficient at epoch time t_E, and $\nu_S(t_E)$ is the topological neighbourhood cut-off coefficient at epoch time t_E.

The synaptic weight of the *j*-th topological connection of the *i*-th node at time *t+1* and epoch t_E, is finally modified as follows:

$$\Delta w_{i,j}(t) = \alpha_S(t_E) \cdot c_{S,i}(t) \cdot [x_j(t) - w_{i,j}(t)]$$

$$w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$$

where $\alpha_S(t_E)$ is the topological learning rate at t_E.

(ii) Temporal learning

On the basis of BMU at time *t-1* and BMU at time *t*, three learning steps are taken:

- temporal connections from BMU at time *t-1* (the *j*-th neuron) to the neighbourhood of BMU at time *t* (the *i*-th neurons) are strengthened:

$$m_{i,j}(t+1) = m_{i,j}(t) + \alpha_T(t_E) \cdot c_{T,i}(t) \cdot [1 - m_{i,j}(t) + \beta_T(t_E)]$$

$$c_{T,i}(t) = e^{-\frac{d_i^2(t)}{2\sigma_T^2(t_E)}}$$

- temporal connections from all neurons but BMU at time *t-1* (the *j*-th neurons) to the neighbourhood of BMU at time *t* (the *i*-th neurons) are depressed as well:

$$m_{i,j}(t+1) = m_{i,j}(t) - \alpha_T(t_E) \cdot [1 - c_{T,i}(t)] \cdot [m_{i,j}(t) + \beta_T(t_E)]$$

$$c_{T,i}(t) = e^{-\frac{d_i^2(t)}{2\sigma_T^2(t_E)}}$$

- temporal connections from BMU at time *t-1* (the j-th neuron) to nodes lying outside the neighbourhood of BMU at time *t* (the *i*-th neurons) are depressed as well:

$$m_{i,j}(t+1) = m_{i,j}(t) - \alpha_T(t_E) \cdot c_{T,i}(t) \cdot [m_{i,j}(t) + \beta_T(t_E)]$$

$$c_{T,i}(t) = \begin{cases} e^{-\frac{d_i^2(t)}{2\sigma_T^2(t_E)}} & if\ d_i(t) \le \nu_T(t_E) \\ 0 & if\ d_i(t) > \nu_T(t_E) \end{cases}$$

(iii) Learning decay

As an epoch ends, an exponential decay process applies to each learning parameter so that the generic parameter *p* at t_E is calculated according to the following equation:

$$p(t_E) = p(0) \cdot e^{-\frac{t_E}{\tau_p}}$$

A complete list of the learning parameters is shown below:

- α_S: learning rate of the topological learning process
- σ_S: shape parameter of the neighbourhood Gaussian function for the topological learning process
- ν_S: cut-off distance of the neighbourhood Gaussian function for the topological learning process
- α_T: learning rate of the temporal learning process
- σ_T: shape parameter of the neighbourhood Gaussian function for the temporal learning

process
- ν_T: cut-off distance of the neighbourhood Gaussian function for the temporal learning process
- β_T: offset of the Hebbian rule within the temporal learning process

(iv) Post processing
At a given epoch t_E, the transition matrix is extracted from the temporal connection weights $m_{i,j}(t_E)$, so that $P_{i,j}(t_E)$ is the probability to have a transition from the *i*-th node to the *j*-th node of the network (i.e., the *j*-th node will be the *BMU* at time *t+1*, given the *i*-th node is the *BMU* at time *t*):

$$P_{i,j} = m_{j,i} \cdot \frac{1}{\sum_{h=1}^{N} m_{h,i}}$$

At the same time the labelling procedure is applied. A label L_i (i.e., an input symbol) is assigned to each node, so that the grapheme-base coding of the *c*-th symbol matches the *i*-th node's space vector best:

$$L_i = \arg\min_c \sqrt{\sum_{j=1}^{D} [x_{c,j}(t) - w_{i,j}(t)]^2} \quad i = 1,...,N$$

A.3 Lexical recall

During the lexical recall task, an activation pattern at time *t* does not die out at time *t+1*, but accrues in the map's short-term buffer. When the whole form is shown, the map's short-term buffer thus retains the integrated activation pattern of all letters of the currently input form. Lexical recall is eventually modeled as the task of restoring the input sequence, by priming the map with the '#' symbol first, followed by the integrated activation pattern. More formally, we define the integrated activation pattern $\hat{Y}\{\hat{y}_1,\ldots,\hat{y}_N\}$ of a word of *k* symbols as the result of choosing

$$\hat{y}_i = \max_{t=2,\ldots k} \{y_i(t)\} \quad i = 1,...,N$$

Lexical recall is thus modeled by the activation function (see Section A.1 above), with

$$y_{S,i}(t) = \begin{cases} \sqrt{D} - \sqrt{\sum_{j=1}^{D} [x_j(t) - w_{i,j}]^2} & t = 1 \\ \hat{y}_i & t = 2,...,k \end{cases}$$

A.4 Parameter configuration

The experiments shown in the present work were performed using the following parameter configuration:
- 40x40 map nodes
- 30 elements in the input vector (orthogonal symbol character coding)
- 100 learning epochs
- learning rates starting from maximum value (i.e. 1.0), exponentially increasing/decaying over epochs (with a time-constant equal to 25 epochs) according to the training error trend
- spatial shape parameter starting from a value so that the Gaussian function has a gain equal to 90% at the maximum cut-off distance, with no decay over epochs
- temporal shape parameter starting from a value so that the Gaussian function has a gain equal to 20% at the maximum cut-off distance, with no decay over epochs
- cut-off distances starting from the maximum distance between two nodes in the map, exponentially increasing/decaying over epochs (with a time-constant equal to 5 epochs) according to the training error trend
- offset of the Hebbian rule within the temporal learning process starting from 0.01), exponentially increasing/decaying over epochs (with a time-constant equal to 25 epochs) according to the training error trend

Developing a lexicon of word families for closely-related languages

Nuria Gala

LIF-CNRS UMR 6166
163 Av. de Luminy case 901 F-13288 Marseille cedex 9
nuria.gala@lif.univ-mrs.fr

Abstract

Lexical resources are of interest in linguistic research and its applications. However, building and enriching them is very time consuming and expensive. In specific fields such as morphology, unsupervised and (semi-)supervised approaches consisting in automatically discovering word structure have gained in popularity in the last few years. While encouraging results have been obtained for a large variety of languages, few resources are currently available. In this paper, we describe a morphological lexicon under development for Romance languages. It is based on an initial seed set of manually identified 2,004 word families in French. Our goal is to map these families on related languages in order to obtain a resource based on family clusters, capable to provide morphological and semantic information on each family crosslingually. Such a resource will be of help in contrastive linguistics and in different NLP and human applications, such as crosslingual information retrieval and interlingual language learning.

1. Introduction

A variety of multilingual lexical resources have been developed by different civilizations ever since the birth of writing, as a result of practical needs (learning, archiving, transmitting linguistic and other kind of knowledge, etc.). The shape and the contents of these resources have evolved significantly over time. Technical revolutions such as printing and computerization have had a profound influence on the way to develop lexicons. From linear presentations of word lists to lexical networks, multilingual lexical resources present interlingual correspondences and often very specific linguistic information.

Obviously, manually building and enriching such resources is very time consuming and expensive. In recent years, collaborative and automatic approaches have emerged as a plausible alternative to build resources in a large-scale perspective thus limiting the time of development. Collaborative multilingual resources such as *Papillon* (Boitet et al., 2002) are based on the principle of sharing contributions, that is, anyone collaborates to enrich the database according to his/her possibilities. While the underlying philosophy is interesting, the results can easily be disappointing, as enriching a resource is a tedious task and in practice few people accept. Hence, it is hard to get the expected volume of contribution[1].

In order to address both shortcomings (manual and contributive), automatic approaches have gained in popularity in NLP, especially when it comes to collecting specific linguistic information. In morphology, different methods exist to automatically acquire information about the internal structure of words (Lavalle and Langlais, 2010): probabilistic approaches which regroup words into paradigms by removing common affixes (*Paramor* (Monson et al., 2007)) or community-based detection in networks (*MorphoNet* (Bernhard, 2010)), unsupervised learning of word structure by decomposition (*Linguistica* (Goldsmith, 2001), *Morfessor* (Creutz and Lagus, 2005)), supervised or semi-supervised methods using formal analogies to identify stems and morphological information (Lepage, 1998), (Hathout, 2008), (Lavalle and Langlais, 2010).

These methods may differ with respect to the kind of result they obtain: word segments, complete morphemic analysis, morphological links between words, etc. Furthermore, as raw text is used for knowledge acquisition, most systems do not make a difference between inflectional or derivational morphology.

The work presented in this paper aims at building cross-linguistic morpho-phonological families. A morpho-phonological family groups lexical units sharing morphological[2] and semantic features. Such a family is usually built around a common stem. Hence, the stem *terre* 'earth', will induce the family made of lexical units such as *terrasse* 'terrace', *terrestre* 'terrestrial', *terrien* 'landowner', etc. All the words in this family share the following semantic components: 'surface', 'ground', 'area', etc. Having translated the stem in closely-related languages and using multilingual corpora and lexica, we will build the corresponding families and compare their organization across languages. Our aim is thus to create a resource presenting word families among closely-related languages and to check whether they can be mapped on each other. The linguistic description provided is strictly synchronical and concerns both derivational morphology (stems and affixes) and morphosemantic links (semantic components within a word family).

This paper is structured as follows. In the next section we provide an overview concerning some existing mono- and multilingual resources by focusing on those containing a morphological description. Section 3 describes first experiments to map our initial resource for French to other Romance languages. We conclude the paper by discussing the achieved results and present some ideas concerning future work.

[1] For a discussion, see (Cristea et al., 2008).

[2] Phonological alternations are possible, i.e. *fleur/flor-* in *fleur* 'flower' and *floraison* 'flowering', *croc/croch-* in *croc* 'hook' and *crochet* 'little hook'.

2. Morphological resources: an overview

Although a significant number of existing NLP lexicons present primarily syntactic or semantic information — subcategorization (Briscoe and Carrol, 1997), concepts as in *WordNet* (Fellbaum, 1998), etc. — an increasing interest in morphology has led over the past few years to the development of morphological lexica. Such resources present a fine-grained and explicit description of the morphological organization of the lexicon. The resources are mainly monolingual, though some multilingual examples can be mentioned.

2.1. Monolingual lexica

For morphology rich languages such as Romance or Slavic languages, monolingual lexica may display morphotactics (ordering of morphemes, derivational morphology) or morphosyntactic information (word forms associated to: a lemma, a part-of-speech tag, inflectional categories, subcategorization patterns, etc.).

The *Digital Dictionary of Catalan Derivational Affixes (DSVC)* (Bernal and DeCesaris, 2008) illustrates a derivational morphology lexicon. It has been created manually and is of limited coverage: about one hundred verbs. *Lefff* (Clement et al., 2004) is an example of morphosyntactic lexicon for French verbs (about 5,000 entries). It has been built automatically by extracting information from large raw corpora and other existing resources.

As for Slavic languages, *Unimorph*[3] is a derivational morphology database with 92,970 Russian words. There is also a morphosyntactic lexicon for Polish (Sagot, 2007) which has been created using the same formalism as the one in *Lefff* through automatic lexical acquisition from corpora. Such a formalism has also been used for other European languages (e.g. Spanish), as well as for less resourced languages such as Kurdish and Persian (Walther and Sagot, 2010).

2.2. Multilingual lexica

Multilingual resources provide the basis for translation, that is, the mapping from one language to the other (Calzolari et al., 1999). Yet this does not always hold for all multilingual morphological resources.

A leading example is CELEX (Baayen et al., 1995), a manually-tagged morphological database for English, Dutch and German. For each language, words are analyzed morphologically and the processes of derivation are made explicit (e.g. 'concern'[V], 'unconcern' ((un)[N—.N], ((concern)[V])[N])[N]). Unfortunately, the morphological information is not explicit crosslinguistically, that is, CELEX is a database for three languages independent one from another.

MuleXFoR[4] (Cartoni and Lefer, 2010) is a morphological database aiming to present word-formation processes in a multilingual environment. Word formation is presented as a set of multilingual rules available by affixes, rules and constructed words (e.g. by the rule *'above (n < a)'*, the following affixes are displayed: *sopre, sovra, super* (Italian), *sur, supra* (French), *supra* (English), along with some words containing such affixes in each language. Word formation processes are thus represented in a multilingual context. Although morphological knowledge was partly automatically acquired from corpora, the coverage of MuleXFoR is limited to one hundred prefixes.

Finally, unsupervised learning of morphologically related words in various languages (English, German, Turkish, Finnish and Arabic) has been the main goal of systems participating to Morpho Challenge 2009[5], e.g. *Morfessor* (Creutz and Lagus, 2005), *Rali-Cof* (Lavalle and Langlais, 2010), *MorphoNet* (Bernhard, 2010), etc. While such competition allows the comparison of different statistical machine learning techniques (in terms of precision and recall), the challenge does not yield any avalaible morphologically annotated resource.

2.3. Remarks

Two general observations can be made at this point: first, very few available resources present morphological links crosslinguistically, and if they do, their coverage is limited; second, morphological processes described by the existing resources mainly focus on word-formation (word construction) conveyed by affixes. To our knowledge, word families — although described in the literature (Bybee, 1985) — have been brought to the forefront only in psycholinguistics to show their impact in lexical decision tasks (Schreuder and Baayen, 1997).

3. Mapping from to other Romance languages

Considering that closely-related languages have a common origin, morphological regularities may be conveyed by means of similar constructions. Our aim is thus to use a manually built morphological lexicon (*Polymots*[6] (Gala et al., 2010), with 2,004 stems and nearly 20,000 derived words for French) and map it to other Romance languages.

3.1. Word families: definition and properties

We consider a family (cluster or paradigm) to be a set of lexical units sharing a formal and a semantic component. Similar words in a lexical cluster share:

- a stem (e.g. *human* in *human*, *humanism, humanist, humanitarian, humanity, humanize, dehumanize*, etc.);
- semantic continuity (all the words in the previous serie are related to the notion of 'bipedal primate mammal').

While in some families there is a continuity of meaning (words sharing a significant number of semantic features, e.g. the *human* family), in others meaning is distributed, i.e. a single and precise meaning is impossible to seize among the lexical units of the cluster, as the words have evolved and the semantic components are widely dispersed.

[3] http://courses.washington.edu/unimorph/

[4] http://www.issco.unige.ch/en/staff/bruno/mulexfor

[5] http://research.ics.tkk.fi/events/morphochallenge2009/

[6] http://polymots.lif.univ-mrs.fr

In such cases, the semantic features of the common stem are to be found among the words in the family (e.g. French *val-* 'glen' includes features such as *geographic area* and *going downhill* and at least one of these notions is to be found in *vallée* 'valley' and *avaler* 'swallow'). Semantic components have been automatically extracted from structured corpora (Gala and Rey, 2009) and are currently being refined with a new machine-readable lexicographic resource (the *Trésor de la Langue Française informatisé*).
Similarly, the size of the families may vary significantly depending on the stem, going from no derivation at all (e.g. French *agrume* 'citrus fruit', *paupière* 'eyelid', etc.) to more than eighty derived words (e.g. *port-* in *apport, emporter, exporter, porable, etc.*). The larger the family the more significant the semantic dispersion; however, the higher the number of analogous word-forms across languages.

3.2. From French to other Romance languages

The French lexicon that we have used to map to the other Romance languages contains 2,004 stems (i.e. 2,004 families). The stems are of two kinds: 87 % (1,741) are lexemes (e.g. *terre* 'ground', *bras* 'arm', etc.); the remaining 13 % (263) are word-forms which do not exist anymore as single tokens (stems in italics, e.g. *bast*ille, *bast*ion, etc.; ap*port*, ex*port*er, etc.; con*verg*ence, di*verg*ent, etc.).
From the initial 1,741 stems, we have conducted a preliminary experience by using a subset of 30 stems (see table 1). To obtain these 30 words, we have selected the most frequent ones from the Greenberg's list, i.e. those having a frequency $f > 0,1$ % on the BNC (70 out of 100). Once these 70 words have been manually translated into French, we have kept those having a frequency $f > 0,01$% in the VocaRef corpora[7] (Table 1). For each word in this list we have automatically acquired their lexical clusters (families) using raw corpora, POS tagged and lemmatized corpora.

grand, dire, voir, homme, venir, donner, savoir, petit, bon, nouveau, personne, femme, entendre, tête, nom, nuit, eau, long, sein, coeur, pierre, humain, mourir, tuer, langue, feu, chemin, bras, sang, oeil
big, say, see, man, come, give, know, small, good, new, person, woman, hear, head, name, night, water, long, breast, heart, stone, human, die, kill, tongue, fire, path, arm, blood, eye

Table 1: List of 30 words from Greenberg's list with $f >$ 0,1 % in the BNC and $f >$ 0,01 % in VocaRef corpora.

3.2.1. Semi-supervised learning using raw corpora

After having translated the 30 seed words in Spanish, Catalan, Italian, Corsican and Portuguese, we extracted all the words from different raw corpora[8] containing every single stem (e.g. in Catalan, *branca*da, *branca*l, *branca*tge, etc. contain the stem *branca* 'branch'). We have refined such first loop with three variants. First, if the stem ends in a vowel, we have retrieved all the forms containing the stem minus the final vowel. We did this in order to address the problem of vocal alternations (e.g. Italian *nome* 'noun, name' / *nominare* 'nominate'). Second, if the stem ends in a voiceless velar plosive (/k/) or a voiceless alveolar fricative (/s/) we had a look at all possible graphical variants: e.g. for the latter, Portuguese and Catalan ç / c (*cabeça / cabecear* 'head / to head', *braç / bracet* 'arm / little arm'). Finally, we considered alternations for diphthongs in Spanish: /e/ with /je/ (*pedrería / piedra* 'jewels / stone'), /o/ with /we/ (*novedad / nuevo* 'novelty / new').
The words obtained were then manually validated via a monolingual dictionary for each one of the respective languages. This allowed us (1) to capture words absent from the corpora and (2) to eliminate candidates wrongly retrieved because we had used only formal analogies without taking into account any semantic information (e.g. in Corsican, the stem *testa* 'head' yields *testa*tu 'stubborn' and in*testa*tura 'header', while cun*testa* 'answer' and *testa*mentu 'will' do not show up in the expected 'head' cluster, eventhough they present the same graphical form). Table 2 shows several members crosslingually gathered for the same family.

FR	oeil	oeillade, oeillard, oeillère, oeillet...
CA	ull	ullada, ullera, ullerat, ullerer, ullerol...
CO	ochju	malochju, ochjuculà, ochjutu...
PT	olho	olhar, olhadinha, olheiras, olhudo...
ES	ojo	ojera, ojear, ojal, ojoso, ojuelo...
IT	occhio	occhialàio, occhiàle, occhialìno...

Table 2: Exemples of 'eye' family.

This first experience reveals that using the stem and some morpho-phonological variants allows us to gather a significant number of candidates belonging to the same family. Not surprisingly, the longer the stem, the higher the accuracy (see 4.1).

3.2.2. Semi-supervised learning based on annotated corpora

A second experience has been carried out to map the initial French stems to other closely-related languages. This time we wanted to scale up to a higher number of families with the same heuristics used in the previous test (stems and their morpho-phonological variants). We also wanted to restrict the mapping to two languages (Catalan and Spanish).
The underlying hypothesis for this second experience is the idea that using annotated corpora would increase the accuracy of the results, mainly because of the absence of inflection. This being so, we used a POS tagged and lemmatized corpora extracted from Wikipedia (258,315 lemmas for Catalan, 387,003 for Spanish)[9]. The corpora have been annotated with lemma and part of speech information using the open source library Freeling[10] 2.1.

[7] 234 millions of words from French newspapers *Libération* and *Le Monde*, 1995-1999.

[8] We have extracted corpora from the Web, mainly Wikipedia and newspaper sites.

[9] http://www.lsi.upc.edu/~nlp/wikicorpus/

[10] http://nlp.lsi.upc.edu/freeling/

We used bilingual corpora to automatically extract the translations of our initial 1,741 stems which are lexemes. The corpora used[11] (7,523 entries FR-CA and 25,616 entries ES-CA) allowed us to extract 30 % of the expected trilingual equivalences, that is 473 stems out of the initial 1,741. From such trilingual set of stem equivalences, we have gathered word forms in the Wikipedia corpora containing each stem and its variants for each of the two languages. At the end of the experience, we have obtained 190 families for Spanish (40,2 %) and 77 for Catalan (16,3 %).

4. Preliminary results

4.1. Raw corpora and stem lengths

The results of hand-validating the data obtained for five languages from raw corpora shed light on significant differences related to the length of words. As we have taken a list of very frequent words, the global average length for all languages is 5 characters as shown in Table 3.

Language	Average	$>= 5$	< 5
CA	4,3	43 %	57 %
CO	4,7	57 %	43 %
FR	4,9	53 %	46 %
PT	5,1	67 %	33 %
ES	5,2	80 %	20 %
IT	5,6	80 %	20 %

Table 3: Word lengths.

Taking into account such a threshold (i.e. 5 characters), precision is about 85 % for stems of length greater or equal to 5 and about 15 % for stems with less than 5 characters. The shorter the stem, the higher the number of word-forms collected, only few being members of the expected family. Furthermore, the shorter the stem, the higher the possibility of homonymy (e.g. in Catalan, *nou* 'new / nine / walnut'), hence the higher the probability to collect word-forms valid from a formal point of view, but unacceptable semantically in a given paradigm (e.g. *noucentista* 'related to the beginnings of years 1900', hence related to 'nine' but not to 'new'). It is also noticeable that in raw corpora inflected and compounded word-forms, as well as misspelled words (e.g. **dinousaure* 'dinosaur'), contribute to decrease the precision rate for all languages (we aimed at collecting only well-formed derived lexical units). As for recall, the data has been manually evaluated by comparing the words obtained with a list of entries present in a monolingual dictionary for each language. As we have used relatively small corpora (100,000 to 300,000 words) global recall is about 50 %, again with significant differences among languages and families.

4.2. Bilingual corpora and analogies among languages

The availability and the size of the resources is crucial for semi-supervised acquisition of information. In our experience, only 30 % of the stems have a correspondence in the three languages. Bigger bilingual corpora is thus necessary to scale up automatically to our initial 2,004 stems as well as to map from French to other languages. At this stage, the comparison of the 473 stems among French, Spanish and Catalan, already gives us some insights concerning the relative linguistic distance of these three languages (cognates, see table 4). Our aim is to consider cognates among families and not only among individual words.

Analogies		example FR CA ES
FR CA ES	69,98 %	*ouvert obert abierto* (open)
FR CA -	5,29 %	*pleur plor llanto* (cry)
- CA ES	16,91 %	*besoin necessitat necesidad* (need)
FR - ES	1,48 %	*corne banya cuerno* (horn)
- - -	6,34 %	*creux buit hueco* (hollow)
	100 %	

Table 4: Lexical closeness among FR CA ES.

About 70 % of the stems are analogous in the three languages, i.e. lexical items share the same form; Catalan and Spanish are closely-related in 86,89 %, Catalan and French in 75,27 % and French and Spanish in 71,46 %.

4.3. Lemmatized corpora and family clusters

The use of a large coverage corpus has enabled us to obtain family clusters very quickly: we have gathered 5,999 word-forms being part of 190 families in Spanish and 1,561 word-forms for 77 families in Catalan. We have conducted an evaluation on 40 stems (40 different families) with 618 words in Spanish and 428 words in Catalan. The average precision is 62,71 % for Spanish and 64,42 % for Catalan, but the results are very heterogeneous among the families. Yet for some families precision is very high (in some cases, 100 %, i.e. all the acquired words belong to the family, see table 5). However, in other families, precision is very low (the word-forms obtained do not belong to the cluster) mainly for reasons of homography (e.g. in Catalan, the string *tendre* 'tender' can be found at the end of a significant number of verbs *abstendre, desentendre, entendre, estendre, etc.*). Some drawbacks come also from the corpora itself: words in other languages, misspellings and errors in tokenization and lemmatization. With the heuristics employed (stems and morpho-phonological variants), we also capture all the existing compounds for a given stem. We are thus considering whether to include them into the resource or to limit it to derivation strictly. Recall is under evaluation using monolingual dictionaries for both languages.

abrigo	abrigo, abrigado, abrigador, abrigamiento, abrigar, desabrigar, desabrigado, desabrigo
aceituna	aceituna, aceitunado, aceitunero, aceitunillo
chocolate	chocolate, achocolatado, chocolatada, chocolatería, chocolatero, chocolatina

Table 5: Family clusters for Spanish.

[11] `http://sourceforge.net/projects/apertium`

5. Conclusion

In this paper, we have presented some initial steps in developing a lexical resource for word families across closely-related languages. The lexicon is constructed from an initial set of stems, identified and manually validated for French. The approach relies on automatically acquiring information from corpora and it reveals that using partially annotated corpora (lemmatized corpora) leads to better results provided that morphophonological properties of each language (e.g. diphthongs in Spanish, consonant and vocalic alternations) are taken into account. We have thus automatically acquired lexical clusters with equivalences in closely-related languages. Word families are connected in order to allow crosslingual access of lexical items via morphological and/or semantic criteria.

Such a lexicon is under development at the time of writing this paper: scaling up to the initial 2,004 stems will be carried out soon for Catalan and Spanish (with bigger annotated corpora and lexica available).

As for future work, we plan to extend the resource to the remaining Romance languages[12]. Furthermore, automatic acquisition of morphological information on analogical series of words (words containing the same affixes, (Hathout, 2008)) is also foreseen shortly. The resulting resource will be freely available and we hope it will be of help for many multilingual human and NLP applications.

6. Aknowledgements

The author would like to thank M. Zock as well as the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

7. References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The celex lexical database release 2. CD-ROM. Linguistic Data Consortium, Univ. of Pensilvania, USA.

E. Bernal and J. DeCesaris. 2008. A digital dictionary of catalan derivational affixes. In *Proceedings of Euralex 2008*, Barcelone, Spain.

D. Bernhard. 2010. Morphonet: Exploring the use of community structure for unsupervised morpheme analysis. In *Multilingual Information Access Evaluation. 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009. Revised Selected Papers.*, volume 1, pages x–x. Springer.

C. Boitet, M. Mangeot, and G. Serasset. 2002. The papillon project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In N. Ide G. Wilcock and L., editors, *Proceedings of on Natural Language Processing and XML, COLING Workshop*, pages 9–15, Taipei, Taiwan.

T. Briscoe and J. Carrol. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Whashington, DC.

[12]Italian, Corsican, Portuguese, Galician and Romanian, possibly Sardinian — the oldest Romance language — provided that enough data is available, i.e. machine-readable annotated corpora and monolingual dictionaries).

J. L. Bybee. 1985. *Morphology. A study of the relation between meaning and form. Typological studies in Language*. Benjamins, Amsterdam.

N. Calzolari, K. Choukri, C. Fellbaum, E. Hovy, and N. Ide. 1999. Multilingual ressources. *Multilingual Information Management : current levels and future abilities. Report commissioned by the US National Science Foundation and the European Commission's Language Engineering Office*.

B. Cartoni and M. A. Lefer. 2010. The mulexfor database: Representing word-formation processes in a multilingual lexicographic environment. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta.

L. Clement, B. Sagot, and B. Lang. 2004. Morphology based automatic acquisition of large-coverga lexica. In *Proceedings of the Fourth conference on International Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal.

M. Creutz and K. Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0, publications in computer and information science. Technical Report A81, Helsinki University of Technology.

D. Cristea, C. Forascu, M. Raschip, and M. Zock. 2008. How to evaluate and raise the quality in a collaborative lexicographic approach. In *Proceedings of the Sixth conference on International Language Resources and Evaluation (LREC'08)*, Marrakech.

C. Fellbaum. 1998. Wordnet: an electronic lexical database. Technical report, MIT Press, Cambridge, MA.

N. Gala and V. Rey. 2009. Acquiring semantics from structured corpora to enrich an existing lexicon. In *Electronic lexicography in the 21st century: new applications for new users (eLEX-2009)*, Louvain-la-Neuve, Belgium.

N. Gala, V. Rey, and M. Zock. 2010. A tool for linking stems and conceptual fragments to enhance word access. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27:153–198.

N. Hathout. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *3rd Textgraphs workshop*, pages 1–8, Manchester, UK.

J. F. Lavalle and Ph. Langlais. 2010. Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques. In *Proc. Traitement Automatique des Langues Naturelles (TALN-10)*, pages x–x, Montreal.

Y. Lepage. 1998. Solving analogies on words: an algorithm. In *Proceedings of the 17th international conference on Computational linguistics (COLING)*, pages 728–735, Montreal.

C. Monson, J. Carbonell, A. Lavie, and L. Levin. 2007. Paramor: minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of*

9th SIGMORPHON Workshop, pages 117–125, Prague, Czech Republic.

B. Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for polish. In *Proceedings of the 3rd Language and Technology Conference*, pages 42–427, Pozna, Poland.

R. Schreuder and R. H. Baayen. 1997. How complex simple words can be. *Journal of Memory and Language*, 53:496–512.

G. Walther and B. Sagot. 2010. Developing a large-scale lexicon for a less-resourced language : General methodology and preliminary experiments on sorani kurdish. In *Proceedings of the 7th SaLT-MiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, Valetta, Malta.

Bilingual lexicon extraction from comparable corpora: A comparative study

Nikola Ljubešić[1], Darja Fišer[2], Špela Vintar[2], Senja Pollak[2]
[1]Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000, Croatia
[2]Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000, Slovenia
E-mail: nljubesi@ffzg.hr, darja.fiser@ff.uni-lj.si, spela.vintar@ff.uni-lj.si, senja.pollak@ff.uni-lj.si

Abstract

This paper presents a comparative study of the impact of the key parameters for bilingual lexicon extraction for nouns from comparable corpora. The parameters we analyzed are: corpus size and comparability, dictionary size and type, feature selection for context vectors and window size, and association and similarity measures. Evaluation against the gold standard shows that window size of 7 with encoded position yields best results. The consistently best-performing association and similarity measures are Jensen-Shannon divergence with log-likelihood. We have shown that very good results can be achieved with small-sized but purpose-built seed lexicons and that problems arising from dissimilarities between the source and the target corpus can be compensated with their sufficient size.

1. Introduction

Bilingual lexica are the key component of all cross-lingual NLP applications and their compilation remains a major bottleneck in computational linguistics. Automatic extraction of translation equivalents from parallel texts has been shown extremely successful (e.g. Och and Ney, 2000; Tiedemann, 2005) but such a scenario is not feasible for all language pairs or domains because for many of them ready-made parallel corpora do not exist and their compilation is slow and expensive. This is why an alternative approach has been increasingly explored in the past decade that relies on texts in two languages which are not parallel but nevertheless share several parameters, such as topic, time of publication and communicative goal (Fung, 1998; Rapp, 1999). Compilation of such comparable corpora is much easier, especially since the availability of rich web data (Xiao & McEnery 2006).

In this paper we describe a set of experiments that serve to systematically determine the impact of the most important parameters for bilingual lexicon extraction from comparable corpora. The parameters we test and analyze are: the size and level of comparability of the corpus used for bilingual lexicon extraction; the type and size of the dictionary used to translate context vectors; the kind of features used to build context vectors and the amount of context that was taken into account; and, last but not least, the association and similarity measures used to compare the vectors across languages. The main contribution of this paper is a systematic comparison of various parameters that can serve as highly valuable guidelines on the collection of corpora and lexica for similar tasks.

The paper is structured as follows: in the next section we give an overview of previous work relevant for our research, Section 3 contains a description of the resources used and the steps taken in the experiment, in Section 4 we present the results of the evaluation of our approach and a discussion after which we conclude the paper with final remarks and ideas for future work.

2. Related work

For the task of bilingual lexicon extraction, parallel corpora provide very good results. However, the availability of parallel corpora is limited to certain language pairs and domains. Therefore, two main lines of research are proposed. The first one aims at bilingual lexicon extraction from comparable (non-parallel) corpora and the second one focuses on using the web to automatically construct parallel corpora (e.g. Fung et al., 2010). Our research falls in the first category.

The seminal papers in bilingual lexicon constructions are Fung (1998) and Rapp (1999) who proposed similar approaches that are based on the word co-occurrence hypothesis. Their main assumption is that the term and its translation share similar contexts. More recent adaptations of these approaches differ in the selection of methods at different stages.

Translation of vectors. At this stage, most researchers use machine-readable dictionaries. Some authors decide to prune out polysemous words in order to exclude semantic noise. Koehn and Knight (2002) build the initial seed dictionary automatically, based on identical spelling features. Cognate detection is used in a similar way by Saralegi et al. (2008), based on longest common subsequence ratio. Déjean et al. (2005), on the other hand, use a bilingual thesaurus instead of a bilingual lexicon.

Context representation. For selecting the representation of a word's context, approaches differ mainly whether they look at a simple co-occurrence window of a certain size or decide to include some syntactic information as well. For example, Otero (2007) proposes binary dependences previously extracted from parallel corpus, while Yu and Tsujii (2009) use dependency parsers and Marsi and Krahmer use (2010) syntactic trees. Instead of context windows, Shao and Ng (2004) use language models.

Building feature vectors. The words in co-occurrence vectors can be represented as binary features, by term frequency or weighted by different association measures, such as TF-IDF (Fung, 1998), PMI (Shezaf and Rappoport, 2010) and, one of the most popular, the log-likelihood score. Others also investigate weighting co-occurrence terms differently if they appear closer to or further from the nucleus word in the context (e.g. Saralegi et al., 2008).

Selection of translation candidates. For ranking candidate translations, different vector similarity measures have been investigated. Rapp (1999) applies city-block metric, while cosine similarity (Fung, 1998) and Dice (Otero, 2007) seem to provide the best results. In addition, some approaches include re-ranking of translation candidates based on cognates detection (e.g. Saralegi, et al. 2008; Shao and Ng, 2004).

3. Experimental setup

In this section we give a detailed account of the experiments we conducted. In order to gain insight into the impact of the most important parameters for bilingual lexicon extraction, we ran a set of experiments in which we adjusted corpus size and the level of comparability of the texts between the languages. Next, we tested the translation of features in context vectors with three dictionaries of different type and size. Third, we tried out several settings of how to build context vectors and which association measure to use and finally, we tested different similarity measures to rank the translation candidates.

Although the parameters change in each run of the experiment, the basic algorithm for finding translation equivalents in comparable corpora is always the same:

(1) build context vectors for all unknown words in the source language and translate the vectors with a seed dictionary;
(2) build context vectors for all candidate translations words in the target language;
(3) compute the similarity for all translated source vectors and target vectors and rank translation candidates according to this score.

3.1 Corpora

Because it was our aim to analyze the impact of the size and comparability level of the corpus used to extract translation equivalents on the quality of the results we decided to use the English-Slovene part of the JRC-Acquis corpus (Steinberger et al., 2006). This is a 20-million-word parallel corpus of legislative texts, which we POS-tagged, lemmatized and filtered out punctuation and function words before we broke it into non-parallel corpora of different sizes and degrees of comparability.

We first took the English part of the corpus and sliced it into 10 equally-sized slices in chronological order, so that the first slice contained the oldest texts in the corpus and the last slice the most recent ones.

We then compared these slices with one another by computing the Spearman rank correlation coefficient (Kilgarriff, 2001) which compares the ranks of *n* most frequent words in each slice of the corpus. Such a comparison shows that slices from the same chronological period are more similar than those from different periods (e.g. the neighboring slices 3 and 4 are much more similar than the distant slices 2 and 9, see Table 1).

Now that we knew how similar or dissimilar these slices were, we were able to build several comparable corpora by taking the English part of the corpus for some slices and the Slovene part that corresponded to the other slices, making sure there was no overlaps between the slices used for one and the other language. In this way we built two sets of subcorpora; the first set consisted of subcorpora that contained slices with a high Spearman co-efficient, i.e. were highly comparable (called 'easy1-5' corpora), and the other set consisted of subcorpora populated with slices that had a low Spearman co-efficient, i.e. were not very comparable (called 'hard1-5' corpora). These two sets of subcorpora with very different levels of comparability were used to study the impact of corpora comparability on the quality of bilingual lexicon extraction.

Both sets of subcorpora consisted of 5 subcorpora, the smallest one containing a single slice per language (approx. 1.6 million content words) and the largest one 5 slices per language (approx. 8 million content words). The differently sized subcorpora were used to establish what is the smallest possible size of a comparable corpus that could still be used efficiently for finding translation equivalents.

High comparability ('easy1-5' corpora)			
Size	Slo slices	Eng slices	ρ
1.6	s3	s4	0.92
3.2	s1+s3	s2+s4	0.93
4.8	s1+s3+s5	s2+s4+s6	0.95
6.4	s1+s3+s5+s7	s2+s4+s6+s8	0.95
8	s1+s3+s5+s7+s9	s2+s4+s6+s8+s10	0.96
Low comparability ('hard1-5' corpora)			
Size	Slo slices	Eng slices	ρ
1.6	s2	s9	0.50
3.2	s1+s2	s9+s10	0.52
4.8	s1+s2+s3	s8+s9+s10	0.59
6.4	s1+s2+s3+s4	s7+s8+s9+s10	0.66
8	s1+s2+s3+s4+s5	s6+s7+s8+s9+s10	0.74

Table 1: Sets of subcorpora used in our experiment.

3.2 Dictionaries

In order to be able to compare vectors in different languages, a seed dictionary is needed to translate features in source context vectors. We tested our approach on three different dictionaries: a general large-sized bilingual dictionary (Grad), a medium-sized Wiktionary that covers basic vocabulary (Wiki), and a small domain-specific lexicon that was extracted from a word-aligned parallel corpus from the same domain (Acquis).

Only content-word dictionary entries were taken into account. No multi-word entries were considered either. And, since we do not yet deal with polysemy at this stage of our research, we only extracted the first sense for each dictionary entry. The seed dictionaries we obtained in this way contained from 2.800 entries (Acquis) to 6.600 entries (Wiki) and 42.700 entries (Grad).

A comparison of the extracted seed dictionaries with the JRC-Acquis corpus shows that even though the Grad dictionary is four times larger than the Acquis lexicon, the token overlap ratio is almost the same (81% vs. 78%). On the other hand, Wiktionary contains a similar amount of entries but they are not very relevant for the corpus in question (78% vs. 41%). We would like to see in our experiments whether reasonable results can be achieved with a small-sized lexicon with good coverage of the corpus vocabulary, so that large dictionaries which are difficult to obtain are no longer required.

	Types		Tokens	
Dict.	Overlap	Ratio	Overlap	Ratio
Grad	11,191	13.82%	5,634,190	81.73%
Wiki	3,122	3.86%	2,831,234	41.07%
Acquis	2,544	3.14%	5,401,254	78.35%

Table 2: A comparison of vocabulary coverage between the three dictionaries and the JRC-Acquis corpus.

3.3 Building and comparing context vectors

In this experiment we limited the task of extracting translation equivalents to nouns only, so we built context vectors for all those nouns that appear in the corpus at least 100 times and have at least 200 features (content words) in their context. We tested different window sizes (5, 7 and 9 lemmas). We compared two settings for feature selection: plain co-occurrence counts (i.e. bag-of-words approach) vs. included information on the position in which a context word appeared (e.g. L3-L2-L1-target_word-R1-R2-R3). With these settings, we extracted 1,105 vectors from the smallest subcorpus up to 2,494 vectors from the largest one.

In this way, we built vectors for all nouns in the source language and for all nouns in the target language. We tested four different association measures to represent features in the vector: relative frequency, pointwise mutual information (PMI), TF-IDF and log-likelihood (LL). Three variations of TF-IDF were taken into consideration: TF-IDF as defined in the information retrieval community (Spärck Jones, 1972), TF-IDF as defined in (Fung, 1998) and Okapi BM25 as the improved baseline in information retrieval (Robertson, 1994). Since none of the variations showed any significant difference, we disregarded the latter two.

Next, we translated words that appeared as features in the source context vector with a seed dictionary (see Section 3.2). If a feature word was not found in the dictionary, it was discarded from the context vector.

As the final step, the translated source context vector was compared to all target context vectors and the translation candidates were ranked according to their similarity score. The similarity measures we explored are: Manhattan and Euclidean distance, Jaccard and Dice indices adapted to non-binary values (Grefenstette, 1994), Tanimoto index (Tanimoto, 1957), cosine similarity and Jensen-Shannon divergence (Lin, 1991).

4. Evaluation

4.1 Automatic evaluation

Evaluation of the results was performed against a gold standard lexicon that was obtained from automatic word-alignment of a parallel corpus from the same domain. In the gold standard, there are several possible translations for the same source word, and we consider any of the variations as an equally suitable translation. The gold standard contains at least one translation for 1,000 source words.

Below we present the results of three experiments that best demonstrate the performance and impact of the key parameters for bilingual lexicon extraction from comparable corpora that we were testing in this research. The evaluation measure used throughout this research is mean reciprocal rank (Vorhees, 2001) on first ten candidates.

We start with the results for the largest subcorpus with a low comparability score (the hard5 subcorpus). The best-performing features for building context vectors turned out to be window size of 7 with encoded position of context words. The best-performing seed dictionary for translating vectors was the Acquis dictionary which was obtained from a small domain-specific word-aligned parallel corpus.

The measure that underperformed drastically on a regular basis under this setting was the Euclidean distance and was therefore removed from the rest of the experiments. Additionally, Dice gave consistently identical candidate lists as Jaccard and was therefore removed from the experiments as well.

The mean reciprocal rank scores for the described measures are given in Table 3. The best-performing combination is Jensen-Shannon divergence with log-likelihood, followed by Jaccard with log-likelihood and TF-IDF.

	relfreq	pmi	tfidf	ll
manh	0.07	0.11	0.15	0.04
jacc	0.70	0.62	0.74	0.74
tanim	0.57	0.49	0.60	0.43
cos	0.60	0.46	0.61	0.44
jenshan	0.68	0.51	0.69	0.78

Table 3: Evaluation of the results for different association and similarity measures on the hard5 subcorpus.

To get a better insight into the relationship between specific similarity measures and association measures, a series of visualizations is given. First, different similarity measures are compared on a boxplot in Figure 1. The variation in the data comes from using different association measures.

Manhattan is obviously overall the weakest similarity measure for this task while Tanimoto and cosine are regularly outperformed by Jaccard and Jensen-Shannon. Jaccard has more consistent results and could be considered the similarity measure of choice if one disregards the difference in association measures.

Additionally, different association measures are compared in the boxplot in Figure 2. Here the source of variation is different results obtained by different similarity measures. Pointwise mutual information obviously underperforms on a regular basis. Relative frequency, TF-IDF and log-likelihood obtain similar results. The variance in log-likelihood is much higher than in the other two association measures which shows its obvious sensitivity to different similarity measures.

In Figure 3 the same association measures are shown, but only for the two best performing similarity measures: Jaccard and Jensen-Shannon. Here the difference between the three best performing association measures becomes clearer. Log-likelihood is the best performing measure, whilst the second best is TF-IDF. The reason for relative frequency to perform that well in our opinion is the fact that the co-occurrence vectors are built from content words only and association measures do not play such an important role as would be if feature selection was less prohibitive.

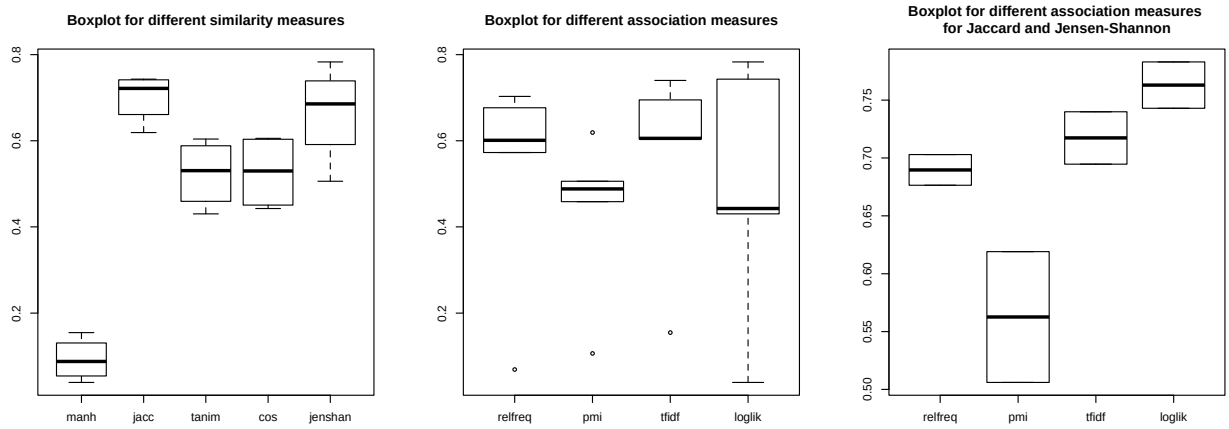
To analyze the consistency of the results, another two experiments were performed under different settings. This time, the smallest (easy1) and the largest (easy5) subcorpora with high comparability scores were used to obtain translation equivalents. These are, as stated before, built from more similar documents than the large, less comparable subcorpus (hard5). Additionally, the easy1 is five times smaller than the easy5 and hard5. In these two experiments, the Grad seed dictionary was used in the vector translation process as opposed to the prior experiment where the Acquis lexicon was used. The Pearson correlation coefficients between the results on hard5 on one side and easy1 and easy5 on the other side are computed. The results are given in Table 4.

	easy1	easy5
all values	0.975	0.982
association measures	0.912	0.957
similarity measures	0.997	0.999

Table 4. Correlation between the results on corpora easy1 easy5 with dict-grad and hard5 with dict-acquis.

The results show a high correlation between all results regardless of the resources and parameters used. When calculating the correlation of different association measure averages, the correlation decreases. On the contrary, when calculating the correlation between results on similarity measure averages, the correlation increases. These results show that specific similarity measures in general have more consistent results regardless of the experiment setting whereas association measures tend to show less consistency. We can conclude that the results of experiments with different settings are highly consistent with association measures being the cause for small variation.

The last experiment we wish to discuss here included different corpus sizes and degrees of comparability. As can be seen in Figure 4, the level of comparability of the corpora plays a major role in the quality of the extracted translation lexicon, especially when very little data is used. However, the size of the corpus is only significant with less comparable corpora. This is a very important finding because corpora with lower degrees of comparability are a much more likely scenario than nearly parallel ones, and it is encouraging to see that by simply increasing their size we can achieve results that are competitive with those obtained from nearly parallel corpora. It must be noted here that since we are using slices of a parallel corpus in this experiment, the level of comparability inevitably increases with corpus size, which is why a similar experiment should be conducted on real comparable corpora in order to confirm our findings in this research.



Figures 1-3: Visualization of the relationships between association and similarity measures regarding the mean reciprocal rank.

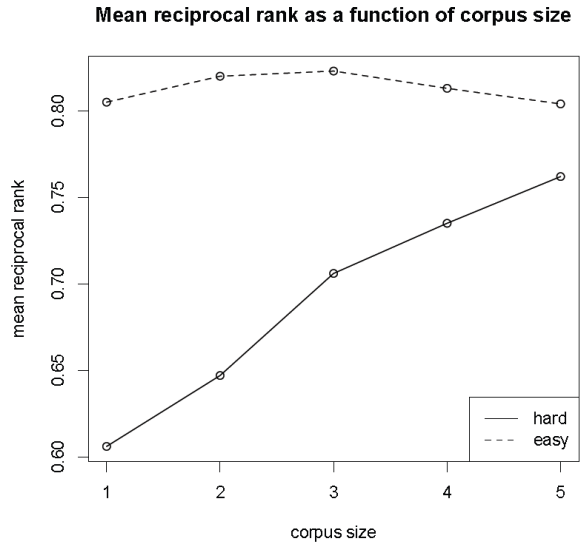


Figure 4. The impact on corpus size and comparability level.

4.2 Manual evaluation

For a more qualitative manual evaluation we selected 100 random source words from the hard5 corpus for which at least one translation candidate was generated, and examined the top ten translation equivalents for each word proposed by our system using the best-performing parameters. In 81 cases the first proposed equivalent matched at least one of the equivalents specified in the gold standard, whereby quite often the list of the extracted equivalents contained all the matches from the gold standard. In 4 cases where the first translation did not match the gold standard we saw that the proposed translation was in fact correct and that the gold standard could have been amended, for example (the correct equivalents are marked in bold):

source word: **integration**

gold standard: povezovanje, vključevanje

proposed equivalents (highest- to lowest-ranking):

integracija	**1.42** (missing in gold standard)
vključevanje	**1.56** (found in gold standard)
povezovanje	**1.59** (found in gold standard)
skupnost	1.64
dialog	1.65
razvoj	1.65
kohezija	1.66
partner	1.66
razsežnost	1.68
sodelovanje	1.69

In 14 cases the correct equivalent was not ranked first and these are the cases we plan to focus on in our future work; we believe that reranking methods applied at the post-processing stage could yet improve these results.

5. Conclusion

In this paper we described a set of experiments we conducted to gain more insight into what really matters in bilingual lexicon extraction for nouns from comparable corpora. The results show that window size of 7 with encoded position of context words are best settings for building context vectors. Small-sized domain specific lexicons that have good coverage of the vocabulary in the corpus can already achieve satisfactory results. This finding justifies the following research scenario as both feasible and efficient: first, a small parallel corpus in the relevant domain is compiled and word-aligned so that a seed lexicon is obtained, and then a much larger comparable corpus in the same domain is used for an extensive extraction of translation equivalents based on the seed lexicon.

What is more, we were able to show that a good combination of an association and similarity measure plays a much bigger role than feature selection or window size. The best-performing combination of association and similarity measures was consistently Jensen-Shannon divergence and log-likelihood. It is interesting to note that while log-likelihood is one of the most popular and best-performing similarity measures in the related work, Jensen-Shannon, which in our experiments outperforms the most popular cosine similarity measure and Dice coefficient, is on the other hand not used as an association measure in any related work we studied. A comparison of corpora of different sizes and degrees of comparability showed that for reasonable results, corpora do not necessarily need to be very similar since the lack of comparability can be compensated to a certain extent with a larger size.

In the future, we wish to test the approach on different corpora, domains and language pairs. In addition, we plan to look at various possibilities to rerank the translation candidates by taking into account cognates and named entities. We also wish to extend our work to other parts of speech and address polysemy as well as multi-word expressions.

6. Bibliography

Déjean, H., Gaussier, E., Renders, J.-M. and Sadat, F. (2005). Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2): 111–124.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pp. 1–17.

Fung, P., Prochasson, E. and Shi, S. (2010). Trillions of Comparable Documents. In *Proceedings of the 3rd workshop on Building and Using Comparable Corpora* (BUCC'10), Language Resource and Evaluation Conference (LREC2010), Malta, May 2010, pp. 26–34.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), pp. 97-133.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the workshop on Unsupervised lexical acquisition* (ULA'02) at ACL 2002, Philadelphia, USA, pp. 9–16.

Marsi, E. and Krahmer,E. (2010). Automatic analysis of semantic similarity in comparable text through syntactic tree matching. In *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), pp. 752–760.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China, pp. 440–447.

Otero, P. G. (2007). Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of the Machine Translation Summit* (MTS 2007), pp. 191–198.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics* (ACL '99), pp. 519–526.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M .(1994) Okapi at TREC-3. *In Proceedings of the Third Text REtrieval Conference (TREC 1994).* Gaithersburg, USA.

Saralegi, X., San Vicente, I. and Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of the 1st Workshop on Building and Using Comparable Corpora* (BUCC) at LREC 2008.

Shao, L. and Ng, H. T. (2004). Mining New Word Translations from Comparable Corpora. *In Proceedings of the 20th International Conference on Computational Linguistics* (COLING '04), Geneva, Switzerland.

Shezaf, D. and Rappoport, A. (2010). Bilingual Lexicon Generation Using Non-Aligned Signatures. In Proceedings *of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010), Uppsala, Sweden, pp. 98–107.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1): 11–21.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), pp. 2142–2147.

Tanimoto, T.T. (1957) IBM Internal Report.

Tiedemann, J. (2005). Optimisation of Word Alignment Clues.*Natural Language Engineering*, 11(03), pp. 279–293.

Vorhees, E. M. (2001). *Overview of the TREC-9 Question Answering Track*. In Proceedings of the 9th Text Retrieval Conference (TREC-9) .

Xiao, Z. and McEnery, A. (2006). Collocation, semantic prosody and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27(1), pp. 103–129.

Yu, K. and Tsujii, J. (2009). *Bilingual dictionary extraction from Wikipedia.* In *Proceedings of the 12th Machine Translation Summit* (MTS 2009), Ottawa, Ontario, Canada.

Construction of a French Lexical Network: Methodological Issues

Veronika Lux-Pogodalla[†], **Alain Polguère**[*†]

[†] ATILF CNRS [*] Nancy-Université
44, avenue de la Libération, B.P. 30687
54063 Nancy Cedex
France
{veronika.lux, alain.polguere}@atilf.fr

Abstract

We present a new lexicographic enterprise that aims at producing a *French Lexical Network* or *FLN*. We begin by introducing the project as such and then proceed with a characterization of the FLN: the FLN as a generic lexical model, its network structure and the different types of lexical entities it models. Finally, we focus on two aspects of our lexicographic methodology: the incremental identification of the FLN's wordlist and our editing tool.

1. The French Lexical Network project

We present a lexicographic project that has just been officially launched (early 2011) and whose aim is to build a new type of lexical resource called *French Lexical Network*, hereafter *FLN*.[1] Though the construction of the FLN is conceived as a long-term enterprise, we focus here on the first three-year phase, i.e. the tasks that have actually been planned and funded in the context of a more global R&D project called *RELIEF*. For lack of space and in order to concentrate on the very specific topic of lexical resources' design and construction, we will ignore the application/valorisation aspects of RELIEF and exclusively deal with the FLN itself.

We are fully aware of the fact that, by presenting a lexical resource that is only emerging from the drawing board, we have no tangible "results" to offer as yet. However, we believe that the FLN project is sufficiently specified, both in terms of design of the lexical resource and of lexicographic methodology, to be of interest for the research community—not to mention the importance there is for the FLN team to benefit from early feedback from this community. Additionally, it will appear clearly in what follows that the FLN is not a project that started from scratch, but a project that directly builds on previous research and lexicographic work performed over the last two decades.

The structure of the remainder of the paper is as follows. The main characteristics of the FLN are presented in section 2.: the FLN as a generic lexical model, its network structure and the different types of lexical entities it describes. In section 3., we focus on two aspects of our lexicographic methodology: the incremental identification of the FLN's wordlist and our editing tool.

Before we begin, let's mention that a lexicographic team of around 15 persons is being put together for the initial three-year phase of the FLN project. Lexicographic strategies embedded in our theoretical and methological framework of reference—the Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995; Mel'čuk, 2006)—will serve to extract linguistic information from corpora. However, we will also make extensive use of the *Trésor de la Langue Française informatisé* (Dendien and Pierrel, 2003), hereafter *TLFi*,[2] as a mother lexical database from which we will extract lexicographic information to be reinterpreted and exploited within the FLN.

2. Main characteristics of the FLN

This section offers a three-step characterization of the global structure of the FLN: the FLN as a generic lexical model (2.1.), the lexical network structure of the FLN (2.2.) and the various types of lexical entities this network connects (2.3.).

2.1. Generic lexical model

In a nutshell, the FLN is designed to belong to the *-Net* family of lexical resources, such as WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 2003; Ruppenhofer et al., 2010). In addition to its network structure, that will be examined in section 2.2., it shares two main characteristics with WordNet and FrameNet:

1. it is **not** a dictionary, i.e. it is not a "textual," linear model of the lexicon;
2. it is nevertheless built in a lexicographic way, i.e. manually by a lexicographic team.

Like WordNet and FrameNet, the FLN has been from the onset conceived of as a generic, general purpose lexical database. However, it is possible to derive a wider range of lexical resources from it: lexicons for NLP systems, full-fledged dictionary entries (similar to those of standard dictionaries) and on-line lexical resources for language teaching/learning. For this reason, it is not focusing on a limited set of specific properties of lexical units (such as semantic hierarchical organization of synsets for WordNet or frames controlling the semantics-syntax interface for FrameNet), but adopts a global view of all lexical properties, very much

[1] The French name of the targeted lexical resource is *Réseau Lexical du Français* or *RLF*

[2] *TLF* stands for the original "paper" dictionary and *TLFi* for its electronic on-line version. The TLFi's URL is: `http://atilf.atilf.fr/tlf.htm`.

like dictionaries would do: lexicographic definition, grammatical features, syntactic combinatorics (roughly, subcategorization frames), lexical combinatorics and derivational links. In that respect, the FLN is equivalent to a virtual dictionary (Selva et al., 2003) or, rather, to virtual dictionaries of various macro- and microstructures that can automatically be generated from it.

2.2. Lexical network structure

The FLN's architecture is similar to a lexical system, as presented in Polguère (2009): a huge network of lexical units connected by a broad range of lexical links encoding semantic or combinatorial lexical relations. The bulk of the network structuring is carried out by means of the system of **standard lexical functions** (Mel'čuk, 1996), that allows for a rigorous encoding of lexical paradigmatic links (synonymy, antonymy, conversivity, actant names, etc.) as well as syntagmatic links (collocations controlled by lexical units—their typical intensifiers, support verbs, etc.). Lexical functions have previously been used in the design of other lexical databases (Fontenelle, 1997; Selva et al., 2003); the FLN is drawing mainly from previous work done on the **DiCo** lexical database (Steinlin et al., 2005) in making use of a double encoding of lexical links:

1. formulas based on the formal language of lexical function relations (Kahane and Polguère, 2001);

2. "popularization" of these formulas in the form of paraphrases (in controlled natural language) of the corresponding paradigmatic or syntagmatic link.

For instance, following this approach, the paradigmatic link holding between KILL$_V$ [*These mushrooms can kill you!*] and LETHAL is to be encoded as follows in the lexicographic article for this sense of KILL$_V$ (popularization comes first, followed by the lexical function formula):[3]

[X] that can $\sim$
Able$_1$ *lethal*

The formal encoding allows for various computations on the lexical graph and the popularization allows for the generation of general public lexicographic descriptions (dictionary articles) from the lexical database.[4]
Beside lexical-functional links, the FLN graph will also encode embedding of semantemes (lexical senses) through its formal definitions—see section 3.2.2. below.
The main aim of the FLN network structuring is to build a model of French lexical knowledge that is truly generic and independent of any specific textual (dictionary-like) or hierarchical (ontology-like) organization. It can also be expected that the chosen model, because of its non-textual nature, will be closer to what is generally believed to be the network-like structure of the mental lexicon (Aitchison, 2003). The main originality of the FLN in terms of structuring, when compared to databases of the *-Net* family, is that it proposes a multi-dimensional graph structure for all standard paradigmatic and syntagmatic links; it does not organize lexical information "through the eyes" of just a few selected links, such as hyperonymy or synonymy. To the best our knowledge, such structure has yet to be implemented, at least for the French language.

2.3. Lexical entities that are nodes of the FLN graph

The FLN will be stored as an SQL database, which will implement its network structure as a set of connections between lexical entities of different types. Central to the lexicographic description are **lexical units** proper, which are of two kinds:

1. **Lexemes** are monolexemic lexical units such as Fr. COUP **I.1** [*Il a reçu un **coup** sur la tête en tombant.*][5] or COUP **I.2** [*Le voleur lui a donné un **coup** sur la tête.*].[6] They correspond to so-called *word senses*.

2. **Idioms** are syntagmatic lexical units such as Fr. ⌜COUP DE SOLEIL⌝ 'sunburn' (lit. 'knock of sun').

Only lexemes and idioms are considered in the FLN as full-fledged lexical units, and they are the actual units of lexicographic description. **Vocables**—polysemic words—are modelled as sets of lexical units connected in the graph by a relation of copolysemy.
The FLN will put strong emphasis on phraseology, i.e. on the set phrases of the language, known as **phrasemes**. Following Mel'čuk (1995), three main types of phrasemes are being considered: (full) idioms, linguistic clichés and collocations.
Because they are lexical units, as much as lexemes are, **idioms** will be described by "normal" lexicographic articles, and not embedded in the article of one of the lexemes they formally contain. For instance, ⌜COUP DE SOLEIL⌝ is not to be described as embedded lexical entity in the article for COUP **I.1**, as it is presently the case in standard language French dictionaries such as *Petit Robert* (Rey-Debove and Rey, 2010).
Linguistic clichés, such as Fr. *Après vous !* 'Go ahead!' (lit. 'After you') are the second type of phrasemes that will be accounted for by lexicographic articles. However, because they are not actual lexical units, clichés will not be considered as "entries" in the database and will receive a somewhat simplified description: no actual lexicographic definition (which will be replaced with the specification of the communicational goals of the speaker) and no indication of combinatorial properties.
As for **collocations**—compositional though phraseological expressions (Hausmann, 1979; Benson et al., 1997)—, they will be accounted for in the article for their base by means of syntagmatic lexical functions, following the approach taken in the DiCo (already mentioned in section 2.2.) and other related lexicographic models.

[3] *X* stands here for the first actant of the keyword (KILL$_V$ = 'X kills Y') and $\sim$ for the keyword itself.

[4] For a general public dictionary (manually) generated from the DiCo database, see the *Lexique actif du français* (Mel'čuk and Polguère, 2007).

[5] *He got a **knock** on his head when he fell.*

[6] *The burglar stroke him a **blow** on his head.*

It can be noted that the lexicological principles adopted for the FLN are very much the same as those of the DiCo project, except for two major differences:

1. Each lexical unit is to be semantically described by a complete and formalized lexicographic definition—whereas the DiCo only provides a description of the actancial structure of the unit together with a semantic label (Polguère, 2003; Polguère, To appear).

2. The data structure of the FLN is a true lexical system, i.e. a network of semantic and combinatorial connections between lexical units. The DiCo's lexical links are in reality connecting lexical units to string of characters (lexical forms), pretty much like any standard dictionary.[7]

By reifying the target of lexical links, the FLN will play in the same "formal" league as WordNet or FrameNet—though its initial vocabulary coverage will of course be very small in comparison (see section 3.1. below, on the FLN's coverage).

3. Lexicographic methodology

This section deals with two methodological aspects of the project that we consider crucial and to which particular attention has been paid: the incremental identification of the FLN's "wordlist" (3.1.) and the writing of FLN articles (3.2.3.).

3.1. The FLN's lexical coverage

3.1.1. Incremental identification of the "wordlist"

In the long run, the FLN should cover the bulk of basic contemporary French. This is a gigantic task, that can only be handled through a series of carefully planned successive efforts. As mentioned earlier, this paper deals exclusively with the initial three-year phase. At the end of this first phase, the FLN should possess a "wordlist"—though the term *wordlist* may not be fully relevant in the specific case of a lexical network—of at least 10,000 vocables.[8] How are these vocables selected among the 70 to 80,000 vocables described in a standard commercial dictionary such as *Nouveau Petit Robert*, idioms included?

The FLN is not designed as a dictionary and, therefore, the process of selecting and building the wordlist can be very different from the selection process implemented by lexicographers of "traditional" dictionaries, such as the TLF, our dictionary of reference (see end of section 1. above). Because of publishing constraints—need for regular releasing of fully completed volumes—, the TLF lexicographers had to first define a whole wordlist, proceeding afterwards through it in strict alphabetical order: vocables starting with the letter *A* being described first, those with letter *B* second, etc. Contrary to this, our progression will not be alphabetical. It will proceed through series of important lexical fields of the language: vocables whose basic lexical unit belong to the semantic field of feelings, of relationships, of animals, of tools, etc. This allows us to start with an initial priming wordlist—Fr. *nomenclature d'amorçage*—that will constantly grow during the project, following a logic that will be detailed shortly.

3.1.2. The priming wordlist

How do we determine the priming wordlist that will be the "seed" from which the whole FLN wordlist will grow in the years to come? In the beginning, priority is given to the most basic, common French vocables. To identify them, we made use of four types of sources:

1. well-known lists of "basic French" developed mainly for applications in language teaching; essentially: the 3,500 vocables of the *Français fondamental* (Gougenheim et al., 1967) and the 3,787 vocables of the *Échelle Dubois-Buyse* (Ters et al., 1988);

2. the "Éduscol" vocabulary list of the 1,462 most frequent lemmas found in the 19th and 20th century French literature;[9]

3. the 6,500 vocables wordlist of the *Robert Benjamin* (Collectif Robert, 2009), a very high quality and seasoned pedagogical French dictionary used in primary schools;

4. a vocabulary wordlist of 4,548 lemmas compiled at the Université de Montréal for the Quebec ministry of education (Ministère de l'Éducation du Loisir et du Sport, MELS) using a meticulous and well-specified methodology (Lefrançois et al., 2011).

Through a cross-checking process,[10] we have identified a priming wordlist of 3,739 vocables, which we believe will induce the description of the basic, minimal set of vocables any speaker of the language, any NLP system, etc., should master.

The number of 3,739 may seem arbitrary, and to some extent it is. This, however, is inconsequential for three reasons. First, it can be noted that most studies on vocabulary thresholds for basic language proficiency conclude to vo-

[7] Of course, a lexical link in the DiCo can specify the actual lexical sense that is the target of the link (`coup#I.1` instead of just `coup`). This, however, is only transparent for the human user of the database and no actual connection is implemented at the level of the data structure.

[8] In comparison, the DiCo—which covers a "sample" rather than a "core" French vocabulary—has a wordlist of 395 finalized (status 0) and 145 prefinalized (status 1) vocables, for a total of 1,127 word senses. The DiCo is accessible on-line in two forms: 1) the *DiCouèbe* interface to DiCo's SQL tables (`http://olst.ling.umontreal.ca/dicouebe`) and 2) the *DiCoPop* dictionary pages automatically generated from the SQL tables (`http://olst.ling.umontreal.ca/dicopop`).

[9] This list, compiled at the Institut National de la Langue Française (INaLF), is available from the Éduscol French government website: `http://eduscol.education.fr/`.

[10] For instance, the *Robert Benjamin*'s wordlist contains many vocables that are mainly relevant in the context of primary school education and by no means belong to the minimal core of French vocabulary—QUADRILATÈRE 'quadrilateral$_N$', SORCIER/SORCRIÈRE 'sorcerer'/'sorceress,' etc. Such vocables are not to be included in the priming wordlist.

cabulary sizes that range from 3,000 "word families"[11] for basic use to 9 to 10,000 for advanced proficiency (Hirsh and Nation, 1992; Nation, 2006). Our 3,739 vocables priming wordlist is therefore in the lower bracket, but still in the realm of what can be considered as a reasonable, basic vocabulary. Second, what matters most is that the vocables we have selected do all belong to basic French and none are peripheral elements of the French vocabulary. Third, it is irrelevant whether one, or two, or 36 vocables have been omitted whose inclusion in the priming wordlist vocabulary would be justified. If a vocable is "missing" for whatever reason, and if it truly belongs to basic French, the induction process that we are now about to describe will catch up with it and have it included in the induced wordlist—Fr. *nomenclature induite*.

3.1.3. The induced wordlist

There are three different ways a vocable that is not present in the priming wordlist can be induced from it: 1) its basic lexical unit is a "close" semantic derivative (nominalization, verbalization, etc.) of the basic lexical unit of a priming vocable, 2) it is a very common idiom formally made up of lexemes of the priming wordlist or 3) its various senses are the target of a significant number of lexical links originating from the lexicographic description of units of the priming wordlist.

1) Induced close semantic derivatives A lexical unit L_2 is a semantic derivative of a lexical unit L_1 if it is the target of a paradigmatic lexical-functional link originating from L_1. The semantic derivation relation between these two units may or may not be marked morphologically.
We use the eleven following paradigmatic lexical-functional links to identify what we term the **close semantic derivatives** of a given lexical unit L.

1. **Syn**: exact synonyms of L, e.g. MOVIE → FILM_N;
2. **Anti**: exact antonyms of L, e.g. LEGAL → ILLEGAL;
3. **of opposite sex Syn$_\cap$**: quasi-synonym (more specifically, intersecting synonym) of L that denotes the same individual/animal as L but of the opposite sex, e.g. ACTOR → ACTRESS, DOG → BITCH;
4. **V_0**: verbal conversion of L, e.g. KNOCK_N → KNOCK_V;
5. **S_0**: nominal conversion of L, e.g. KNOCK_V → KNOCK_N;
6. **Adj_0**: adjectival conversion of L, e.g. COAST_N → COASTAL;
7. **Adv_0**: adverbial conversion of L, e.g. SLOW_Adj → SLOWLY;
8. **S_i**: nouns meaning 'i$^{\text{th}}$ actant of L', e.g. DRIVE_V → DRIVER [= **S_2**];
9. **A_i**: adjectives meaning 'that is the i$^{\text{th}}$ actant of L', e.g. HUNGER → HUNGRY [= **A_1**];
10. **Able$_\text{i}$**: adjectives meaning 'that has the ability to be the i$^{\text{th}}$ actant of L', e.g. LOVE_V → LOVABLE [= **Able_2**].
11. strict **Mult**: collective nouns that do include in their definition the meaning of L, e.g. LEAF → FOLIAGE—but SCHOOL [*of fish, shrimps...*] is not induced directly from FISH, as it is too vague.

Notice that the eleven above-mentioned lexical functions are used here in their "narrow sense," described in the glosses that accompany them. For instance, strictly speaking, VICTIM [*of a murder*] is a valid **S_2** for MURDER_N [*by X of Y*], but it should not be considered as being a close semantic derivative because its meaning is much vaguer than 'Y of a murder' (**murderee*).
It is good practice in Explanatory Combinatorial Lexicography to describe a vocable V together with all vocables whose basic lexical unit (basic sense) is a close semantic derivative of the basic lexical unit of V. For instance, MURDER_N should necessarily be lexicographically described together with MURDER_V, MURDEROUS, MURDERER and MURDERESS. In order to adhere to Explanatory Combinatorial methodology, we consider as induced vocables all vocables whose basic lexical unit is a close semantic derivative of the basic lexical unit of a priming vocable. For instance, though PRÉVISION 'prediction' is not in our priming wordlist, it is included into the induced wordlist as it is a close semantic derivative of the priming vocable PRÉVOIR 'predict.' Notice however that, at this stage, only close semantic derivatives that are commonly used and do not belong to specialized vocabularies will be induced. For instance, HASE 'femal hare' is a close semantic derivative of LIÈVRE '(male) hare,' which belongs to the priming wordlist, but it will not be directly induced from it because of its almost technical nature.

2) Induced idioms The priming wordlist is made up of lexemic vocables. Any common idiom that is formally made up of lexemes that belong to the priming wordlist will be systematically included into the induced wordlist. For instance, as COUP, DE and SOLEIL (see section 2.3. above) all belong to the priming wordlist, ⌜COUP DE SOLEIL⌝ 'sunburn' is identified as induced vocable and added to the lexicographic team's in-tray.

3) High degree nodes of the graph In the process of describing vocables of the priming wordlist, lexicographers will be lead to introduce "on the fly" many new nodes in the FLN graph. They correspond to lexical units that are the target of links originating from priming lexical units. Two main types of links have to be considered.
Firstly, any lexical unit used in a lexicographic definition for a priming lexical unit is necessarily the target of a lexical link (of semantic inclusion). If this target is itself a priming lexical unit, nothing needs to be done. If it is not, a minimal entry for it is generated on the fly in order to make the link hold[12] (on FLNs' definitions, see section 3.2.2. below). For instance, if ASTRE—a very basic but not so com-

[11] In P. Nation's terminology, a word family is a word morphological base form plus all its associated inflectional variants and regular morphological derivations.

[12] Of course, it can also be the case that this unit, though not priming, is already present in the lexical graph as a result of an earlier on-the-fly generation.

mon term roughly equivalent to 'celestial body'—is used as a generic component in the definition of the lexeme SOLEIL 'sun,' then ASTRE will be included in the FLN graph, with minimal information (mainly, its part of speech and some illustrative linguistic examples).
Secondly, any lexical unit that is the target of a lexical-functional link originating from the description of a priming lexical unit also has to be inserted on the fly in the FLN graph. For instance, the adjectival unit RETENTISSANT 'resounding' will be inserted in the graph though it is not among the 3,739 units of the priming wordlist because COUP **I.1** 'knock' is a priming lexical unit and RETENTISSANT is one possible **Magn** (= intensifier) for it.[13]
As a result of the strategy of on-the-fly creation of entries for targeted lexical units, the FLN graph will gradually incorporate a large number of roughly sketched nodes that did not belong to the priming wordlist. A statistical analysis of the graph will regularly be performed in order to identify a list of top non-priming nodes of the graph that possess a high degree of connectivity. These nodes define the next batch of vocables to be inserted in the induced wordlist.
Figure 1 visualizes the wordlist expansion via insertion of idioms, close semantic derivatives and targeted units.

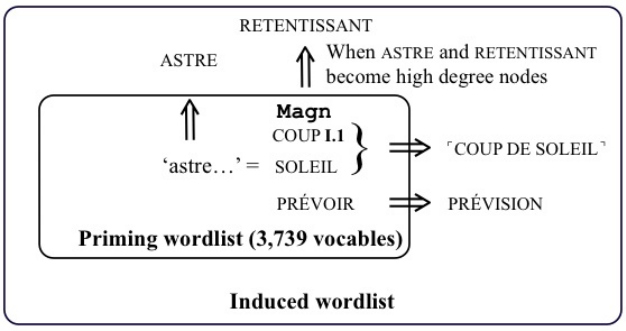


Figure 1: Self induced expansion of the wordlist

As we see, starting from the initial priming wordlist, the FLN will induce its own expansion according to a very simple logic: lexical units that are often referred to by units of the priming wordlist are "important" units, on which lexicographic work should focus. This strategy can be applied indefinitely as a guide to the expansion of the FLN.

3.2. Writing of FLN's articles

3.2.1. In lexicography, size matters

Work previously done on the DiCo database and on other extremely rich and formalized lexicographic models based on Explanatory Combinatorial Lexicology (Mel'čuk et al., 1984 1988 1992 1999) has shown that even skilled lexicographers fail to ensure the coherence of such lexical models when they grow to more than a thousand entries or so. If one wants to use this kind of approach to embark on a major lexicographic project, a rich tailor-made editing environment is required.
FLN articles have to comply to a well-specified structure that could be encoded, in theory, as an XML schema, and enforced through the use of an XML editor. However, there are two aspects of the FLN project that make it impossible to rely on such basic lexicographic tools.
Firstly, building the FLN is a true, large-scale lexicographic enterprise involving the coordinated work of an organized lexicographic team. There is therefore a need to possess an editor that, on top of ensuring the control of the formal validity of the description, will implement a lexicographic production line, with its various tasks (drafting, development, completion with corpus data, revision cycles, etc.) and their logical organization—a workflow management tool system.
Secondly, what really makes the editing of an FLN article complex is the fact that the information it contains has to be stored not as text, but as a database of connected entities forming a lexical graph. We believe that only this type of data structure will ultimately allow us to perform efficient consistency checks and other logical operations on our model of the lexicon. We are particularly interested in the possibility of using the graph structure of the FLN and formal properties of lexical-functional links to implement semi-automatic drafting of vocables based on potential analogies with already existing descriptions—on this, see Jousse (2010, p. 236–257).
Off-the-shelf professional dictionary production softwares such as TLex[14] (de Schryver and de Pauw, 2007) do exist and are used to build major commercial dictionaries. In our case, we chose to work in close collaboration with MVS Publishing Solutions,[15] our partner in the RELIEF project (see section 1.), to tune their Dixit editor for our specific needs. This editor is a component of a software suite mainly used for the publication of daily newspapers. It controls the writing process of newspaper articles (structuring of the article, handling of its editorial cycle and SQL storage of textual as well as non-textual information), data management and automatic generation of printed articles based on predefined layout rules. Thus, it already contains all functionalities one needs in order to perform the writing and, even, publication/dissemination of lexicographic articles.
In the remainder of this section, we will first describe the FLN microstructure the editor has to handle (3.2.2.), then explain the main features of the editor (3.2.3.).

3.2.2. Structure of a lexicographic article

The structure of an FLN article is very similar to that of a DiCo record (Lareau, 2002; Jousse and Polguère, 2005), and an SQL export of the DiCo data is actually being used for tuning the FLN lexicographic editor. As can be seen in Figure 2 below, with the article for ADMIRER **I** 'to admire [someone for something],' an FLN article is divided into six main sections, of which only the second one–Definition–is absent from DiCo records and will therefore be presented in some detail here.

1. **Grammatical features** This section lists features encoding combinatorial properties of the keyword (register, part of speech, inflectional restrictions, etc.).

[13] More precisely, it corresponds to the semi-standard lexical function **in respect to noise Magn**, or $\mathbf{Magn}_{\mathbf{noise}}$.

[14] http://tshwanedje.com/
[15] http://www.mvs.fr/

2. **Definition** In the FLN, each full lexical unit is to be semantically described by means of a paraphrastic definition (which was not the case in the DiCo). Each definition is made up of two components:

   a. The *definiendum* is a description of the actancial structure of the keyword.

   b. The *definiens* (definition proper) is the analytical paraphrase of the keyword's meaning. Prototypically, a definiens is mainly made up of a central component (CC) and one or more peripherical components (PC). Lexicographers annotate the text of the definiens so as to make its internal structure explicite. For example, the definiens in Figure 2 below is encoded in the background as follows:[16]

   ```
   <DEFINIENS label="apprécier">
        <CC>L'individu X apprécie Y pour Z</CC>
        <PC role="intensity"> beaucoup</PC>
        <PC role="cause"> du fait des qualités
            exceptionnelles de Z</PC>
   </DEFINIENS>
   ```

   As indicated in 3.1.3., each lexical item occurring in the definition is connected by a semantic inclusion link to a specific lexical unit—whether priming, induced or pending description—, whose own definition, if it exists, will be subjected to the same formal treatment. Of course, such strategy will make the process of writing a lexicographic definition very slow and, in some respects, tedious. It should be noted, however, that it has the positive effect of forcing lexicographers to proceed very selectively and with economy in writing lexicographic definitions, thus ensuring the production of definitions of greater clarity[17]—see, for instance, the systematic use of a basic defining vocabulary in the definitions of the Longman dictionary (Summers, 2005).

3. **Government pattern** This section describes how the keyword's semantic actants can be expressed as its syntactic dependents. A database of French government patterns will be included in the FLN data structure and valency tables (roughly, subcategorization frames) appearing in a lexicographic article will ultimately be directly imported from this base rather than manually typed by lexicographers.

4. **Lexical functions** This section is the core of the lexical description, as explained in 2.2. Lexical links implemented here will be the main structuring elements of the FLN lexical graph. For lack of space, we cannot enter into the details of the encoding of paradigmatic and syntagmatic links by means of lexical functions. This topic is largely dealt with in the literature on Explanatory Combinatorial Lexicology cited in this paper.

5. **Examples** This section of FLN articles will be much more structured than what can be seen in Figure 2, where only examples imported from the DiCo appear. In an actual FLN article, there will be several types of lexicographic examples, mainly: citations from texts of various genres with exact references—extracted from Frantext[18] and other ATILF in-house corpora—and hand-crafted adaptations of corpus/Internet data.

6. **Phraseology** This last section lists idioms or linguistic clichés that formally contain the keyword. Each enumerated phraseme is linked to the corresponding FLN article.

3.2.3. Designing a lexicographic editor

Recall that the unit of lexicographic description in the FLN is the lexical unit: lexeme ("word" taken in one specific sense) or idiom. Though other lexical entities—such as linguistic clichés (cf. 2.3. above)—may be described by means of lexicographic articles, the editor is essentially providing an interface for lexicographers to describe properties of lexical units.

The lexicographic editor for FLN is currently being prototyped by MVS Publishing Solutions using their Dixit general purpose editor. Figure 2 below is a sample screendump of the editor's interface in its present, very preliminary state. It shows the ADMIRER I entry, based on DiCo data to which a full-fledged lexicographic definition has been added. The purpose of this figure is mainly for the reader to visualize better the type of lexicographic data we are dealing with.

The editing interface helps lexicographers produce descriptions that comply to the microstructure presented above. Practically, their task is closer to filling-in a very complex and structured form than to performing free writing, which is precisely what is required for lexicographic tasks. Moreover, in each section, the editor provides assistance to control compliance to particular constraints on content, ultimately ensuring that the entry is built as a valid subgraph of the global FLN. Depending on the constraints, the level and type of assistance will vary in each section, for instance:

- Normalized content can be directly selected from menus. Text items selected from menus are non-editable text in the article. (They can only be modified through menu selection.)

- Normalized content can also be selected via a form providing filtering features. This is for instance the case with the *Lexical functions* section: there are hundreds of potential lexical function formulas, too many for a single menu. Lexicographers can either indicate some features of the lexical function they are looking

[16] For more information on this approach to formally structuring lexicographic definitions, see Barque et al. (2010).

[17] The rather wordy definition for ADMIRER I in the TLF is *Considérer quelqu'un ou quelque chose avec un sentiment d'étonnement mêlé de plaisir exalté et d'approbation, le plus souvent motivé par la supériorité qu'on lui reconnaît dans divers domaines de la vie intellectuelle, esthétique, morale, etc.* 'To consider someone or something with a feeling of mixed exalted pleasure and approbation, usually motivated by the superiority one acknowledges to him/it in various aspects of life—intellectual, esthetic, moral, etc.'

[18] http://atilf.atilf.fr/frantext.htm

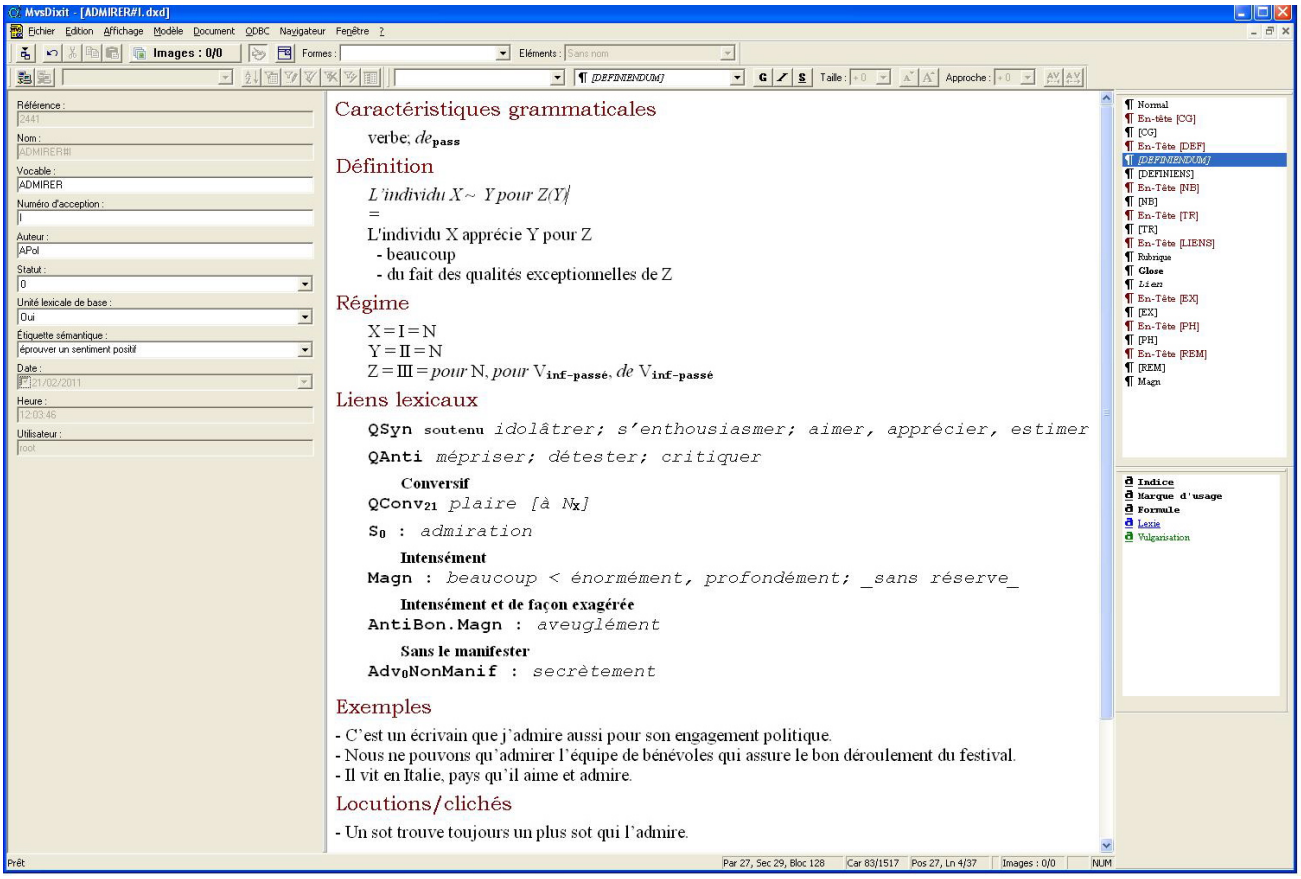


Figure 2: DiCo's data for ADMIRER **I** 'to admire [someone for something]' processed with the FLN editor

for (part of speech of the lexical target, etc.) in a form and get a filtered list of lexical functions in a menu, or they can start typing in the name of the function and get a list of suggestions (through a completion function). Once inserted, lexical function names are non-editable text.

4. Conclusion

As mentioned at the very beginning of this paper, we are presenting a new lexicographic project and it is too early for us, at the time of writing, to be able to draw any conclusion from our theoretical and methodological choices. However, we believe the content, structure and methodological design of the FLN to be original enough to generate interest for anyone concerned with the construction and availability of multi-purpose lexical resources. Of particular relevance is the fact that the FLN is designed as a truly generic database. It targets NLP exploitation—that imposes very strong formal constraints on lexical data—as well as pedagogical exploitation—that shows zero tolerance to error in the modeling of linguistic rules.

Note that the FLN will be made available on the CNRTL website[19] in the course of its growth, both as a source SQL database and via a web-based interface for manual consultation. It is also our intention to later explore the possibility to generate LMF[20] compatible exports of FLN data.

[19] http://www.cnrtl.fr/

[20] Lexical Markup Framework, ISO-24613:2008 (Francopoulo et al., 2009).

Acknowledgements

Many thanks to Sébastien Haton, Jasmina Milićević, Dorota Sikora and two reviewers for WoLeR 2011 for their comments on a preliminary version of this paper. We are very grateful to Pascale Lefrançois (Université de Montréal) and Ophélie Tremblay (Université du Québec à Montréal) for giving us access to their research material, that helped us greatly in the construction of the FLN priming wordlist; we additionally thank Caroline Bégin (MELS) and Hélène Cajolet-Laganière (Université de Sherbrooke) for authorizing the dissemination of the information contained in Lefrançois et al. (2011). The RELIEF project is supported by a grant from the Agence de Mobilisation Économique de Lorraine (AMEL) and Fonds Européen de Développement Régional (FEDER).

5. References

J. Aitchison. 2003. *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell, Oxford UK, 3rd edition.

C. F. Baker, C. J. Fillmore, and B. Cronin. 2003. The Structure of the FrameNet Database. *International Journal of Lexicography*, 16(3):281–296.

L. Barque, A. Nasr, and A. Polguère. 2010. From the Definitions of the *Trésor de la Langue Française* To a Semantic Database of the French Language. In A. Dykstra and T. Schoonheim, editors, *Proceedings of the XIV Euralex*

International Congress, pages 245–252, Leeuwarden, 6–10 July. Fryske Akademy.

M. Benson, E. Benson, and R. Ilson. 1997. *The BBI Dictionary of English Word Combinations*. John Benjamins, Amsterdam/Philadelphia, revised edition.

Collectif Robert. 2009. *Le Robert Benjamin*. Le Robert, Paris.

G.-M. de Schryver and G. de Pauw. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of TshwaneLex. *Lexikos*, 17:226–246.

J. Dendien and J.-M. Pierrel. 2003. Le Trésor de la Langue Française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues (T.a.l.)*, 44(2):11–37.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press., Cambridge, MA.

T. Fontenelle. 1997. *Turning a bilingual dictionary into a lexical-semantic database*. Niemeyer, Tübingen.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2009. Multilingual Resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70.

G. Gougenheim, R. Michéa, P. Rivenc, and A. Sauvageot. 1967. *L'élaboration du français fondamental*. Didier, Paris.

F. J. Hausmann. 1979. Un dictionnaire des collocations est-il possible ? *Travaux de littérature et de linguistique de l'Unviersité de Strasbourg*, XVII(1):187–195.

D. Hirsh and P. Nation. 1992. What Vocabulary Size Is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language*, 8(2):689–696.

A.-L. Jousse and A. Polguère. 2005. *Le DiCo et sa version DiCouèbe. Document descriptif et manuel d'utilisation*. Technical report, Department of Linguistics and Translation, Université de Montréal.

A.-L. Jousse. 2010. *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. Ph.D. thesis, Université de Montréal & Université Paris Diderot (Paris 7), Montreal & Paris.

S. Kahane and A. Polguère. 2001. Formal Foundation of Lexical Functions. In *Proceedings of the Workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation"*, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics, pages 8–15, Toulouse, 7 July 2001.

F. Lareau. 2002. A practical guide to writing *DiCo* entries. In *Proceedings of PAPILLON 2002 International Workshop on Multilingual Lexical Databases*, Tokyo, 16–18 July.

P. Lefrançois, O. Tremblay, and V. Lombard. 2011. Constitution de listes de mots pour l'apprentissage de l'orthographe et du lexique au primaire et au début du secondaire. Research report for the *Ministère de l'Éducation, du Loisir et du Sport du Québec (MELS)*, Université de Montréal, Montreal, 14 February 2011.

I. Mel'čuk and A. Polguère. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Champs linguistiques. De Boeck & Larcier, Brussels.

I. Mel'čuk, A. Clas, and A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Paris/Louvain-la-Neuve.

I. Mel'čuk et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques. Volumes I–IV*. Les Presses de l'Université de Montréal, Montréal.

I. Mel'čuk. 1995. Phrasemes in Language and Phraseology in Linguistics. In Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 167–232. Laurence Erlbaum Associates, Hillsdale, N.J.–Hove, UK.

I. Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins, Amsterdam/Philadelphia.

I. Mel'čuk. 2006. Explanatory Combinatorial Dictionary. In Giandomenico Sica, editor, *Open Problems in Linguistics and Lexicography*, pages 225–355. Polimetrica, Monza.

P. Nation. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, 63(1):59–81, September.

A. Polguère. 2003. Étiquetage sémantique des lexies dans la base de données DiCo. *Traitement Automatique des Langues (T.a.l.)*, 44(2):39–68.

A. Polguère. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43(1):41–55, March. Springer.

A. Polguère. To appear. Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de lexicologie*.

J. Rey-Debove and A. Rey, editors. 2010. *Nouveau Petit Robert*. Le Robert, Paris.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley CA.

T. Selva, S. Verlinde, and J. Binon. 2003. Vers une deuxième génération de dictionnaires électroniques. *Traitement Automatique des Langues (T.A.L.)*, 44(2):177–197.

J. Steinlin, S. Kahane, and A. Polguère. 2005. Compiling a "classical" explanatory combinatorial lexicographic description into a relational database. In *Proceedings of the Second International Conference on the Meaning Text Theory (MTT'2005)*, pages 477–485, Moscow.

D. Summers, editor. 2005. *Longman Dicionary of Contemporary English*. Pearson Longman, Essex, 4th edition.

F. Ters, G. Mayers, and D. Reichenbach. 1988. *L'échelle Dubois-Buyse*. OCDL, Paris.

Different Approaches to Automatic Polarity Annotation at Synset Level

Isa Maks, Piek Vossen

Vu University, Faculty of Arts
De Boelelaan 1105, 1081 HV Amsterdam
E-mail: e.maks@let.vu.nl, p.vossen@let.vu.nl

Abstract

In this paper we explore two approaches for the automatic annotation of polarity (positive, negative and neutral) of adjective synsets in Dutch. Both approaches focus on the creation of a Dutch polarity lexicon at word sense level using wordnet as a lexical resource. The first method is based upon the simple transfer of an English sentiment lexicon (Sentiwordnet 1.0) into Dutch. The second approach regards the use of a wordnet based propagation algorithm with different settings with respect to the quality and length of the seed lists. Results are validated against manually compiled gold standards and compared with results of similar approaches generating polarity lexicons for English.

1. Introduction

The automatic extraction of opinions, emotions, and sentiments in text to support applications such as product, hotel and film review mining, analysis of opinionated text like news, forum posts, and blogs is an active area of research in natural language processing. Many approaches to opinion and sentiment analysis rely on lexicons or lists of words that may be used to express sentiment. Knowing the polarity (positive, negative or neutral) of these words helps a system recognize the positive and negative sentiments in these sentences.
Many subjectivity lexicons are compiled as lists of keywords, rather than word meanings. However, words may have positive, negative and neutral meanings (cf. ex. (1a) and ex. (1b)) which may cause major errors if incorrectly tagged in the applications they are used in.

Ex. (1) wreed (cool, cruel)
(a) een wrede despoot *a cruel tyrant*
(b) ze rijden daar in vet wrede auto's rond *They drive around in really cool cars*

The example shows that the Dutch word *wreed* has two different meanings (properly translated into *cruel* and *cool* , respectively), with opposite (negative and postive) polarity.
Most studies, nowadays, recognize the importance of sentiment scores at meaning level (Esuli and Sebastiani (2006), Andreesvkaia and Bergler (2006), Wiebe and Mihalcea (2006), Su and Markert (2008). Although these approaches are widely used in English, little is known about how they perform at synset level as opposed to word level. More recently, a number of approaches have been tested to build subjectivity lexicons at synset level (Gyamfi et al. (2009); Su et al. (2009)). They focus, however, on subjectivity classification, a task that slighly differs from ours, as it aims at the classification of word senses as subjective or objective.
For Dutch, the only existing polarity lexicon - to our knowledge - is built by Jijkoun and Hofmann (2009). Their approach is, like ours, wordnet based, but produced a list of words (instead of synsets).
In this paper, we focus on the creation of subjectivity lexicons at word sense level using wordnet as a lexical resource where word senses are organized in synsets.
We explore two methods for polarity annotation of Dutch adjective wordnet entries, leaving the nouns and verbs for future work. The first method relies on the transfer of polarity values from an English sentiment lexicon, Sentiwordnet 1.0 (Esuli and Sebastiani (2006)) to the Dutch Wordnet.
The second approach consists of the implementation of a propagation algorithm that starts with a seed list of synsets of known sentiment and sends polarity through the wordnet making use of its lexical relations. Experiments with different seed lists are performed : the General Inquirer word list (Stone et al., 1966) translated into Dutch, and two different manually compiled synset lists following a method that might be used when no manually compiled seed lists exist.

The remainder of this paper is organised as follows. In the next session we briefly discuss the lexical resources and gold standards referred to in this paper. Sections 3, 4 and 5 present the two different approaches to polarity annotations and their results. In Section 6 the results are compared with other studies.

2. Descriptions of Lexical Resources and Gold Standards

2.1 Lexical resources

- **Dutch**

We make use of two lexical resources for Dutch: the Dutch Wordnet and the Dutch Reference Lexicon which both are part of the Cornetto database (Vossen et al. 2008). The two combined resources have different semantic organisations: the Dutch Wordnet has, like the Princeton Wordnet, a synset organisation and the Dutch Reference Lexicon is organised in form-meaning composites or lexical units. The description of the lexical units includes definitions, usage constraints, selectional restrictions, syntactic behaviours, illustrative contexts, etc. Within the Cornetto Database, each synonym in a synset is linked to the corresponding lexical unit of the Dutch Reference Lexicon. Synsets are linked by translation equivalents links to the Princeton Wordnet (versions 2.0 and 3.0); these translation links have been derived automatically and are then manually corrected.
The Cornetto database is semi-automatically compiled and manually corected afterwards. As the manual correction is still in progress , the status of the synsets with regard to the number of lexical relations like hyponyms, near-synonyms, hypernyms and antonyms (LR) and/or translation equivalent links (Equi) may differ. Table 1 presents the statistics of the adjective part of Cornetto. Part ADJ1 consists of 3,616 synsets which have both lexical and translation equivalent relations; Part ADJ2 consists of 2,109 synsets which have translation equivalent relations only; part ADJ3 consists of 733 synsets which have lexical relations only ; and part ADJ4 consists of synsets lacking both lexical and translation equivalent relations.

	Synset	**LR**	**Equi**
ADJ1	3,616	+	+
ADJ2	2,109	-	+
ADJ3	733	+	-
ADJ4	1,440	-	-
Totals	7,898	4,349	5,725

Table 1. Number of Adjective Synsets and Lexical Units in Cornetto (situation 2010)

Because of the different stages of elaboration of the synsets, the two approaches discussed in this paper are relevant for the Dutch wordnet as they may complement each other. Synsets that have translation equivalent links to the English wordnet are covered by the transfer approach and synsets that have lexical relations are covered by the propagation method.

2.2 Gold Standards

- **Dutch**

For the evaluation of the results for Dutch we use the gold standard developed by Maks and Vossen (2010b). The gold standard includes annotations for subjectivity (subjective vs. objective), attitude holder (SpeakerWriter or AgentExperiencer) and polarity (positive/negative/ neutral). Only the latter category will be used in this study. We use the synset level variant of the gold standard which includes 512 synsets (gs-ss-512).
Reported inter-annotator agreement for polarity, is 86.3% with a Cohen kappa (κ) of 0.80. The polarity annotations are distributed as follows: 37% negative, 35% positive, 28 % neutral.
In section 6 we refer to a word level gold standard for Dutch (w-1916) compiled by Jijkoun and Hoffman (2009), which consists of 1916 words annotated by two annotators with positive (50%), negative (29%) and neutral (21%) polarity. Interannotator agreement is 76% (κ=0.62); we use a version where disagreements are abjudicated by a third annotator.

- **English**

For English, the Micro-WNOp corpus (Cerini et al., 2007) is used as a gold standard to evaluate Sentiwordnet. The Micro-WNOp corpus is a – publicly available - list of about 1000 WordNet synsets (285 adjective synsets) annotated with polarity values. The raters manually assigned a duplet of numerical scores to each synset which represent the strength of positivity and negativity, respectively. Thus, a synset could have a non-zero rating on both negativity and positivity. The gold standard does not provide a abjudicated judgment for each synset but the lists with judgments by all different annotators can be downloaded. The gold standard consists of 285 adjective synsets divided into three groups: a common part of 29 adjective synsets with one abjudicated annotation judgment; group 1 consisting of 147 synsets with 2 annotation judgments for each synset and group2 consisting of 138 synsets with 3 annotation judgments for each synset.
For our purposes, we converted the numerical scores to categorical ones (positive , negative and neutral) by assigning 'positive' to synsets where the positive score is larger than the negative score and 'negative' where the negative score is larger than the positive one. The rest of the synsets (i.e. where the positive and negative scores are equal, including zero) is considered 'neutral'. We then derived one judgment for each adjective synset when there was agreement between at least two annotators. The remaining 12 synsets on which all (2 or 3) annotators disagreed were eliminated from the gold standard. Thus,

the final 'simplified and categorical' gold standard which will be called WNO-273 in the remainder of this paper, consists of 273 synsets (78 negative; 70 neutral, 125 positive).

3. Method I: Sentiwordnet translated

The transfer of Sentiwordnet 1.0 to Dutch consists of the copying of the sentiment values from the English synsets to the Dutch synsets through the translation equivalents which exist between the English and the Dutch wordnet. We evaluate the English and the Dutch version and compare the results.

3.1 Method I

Sentiwordnet1.0 (Esuli and Sebastiani (2006)) is a resource with automatically determined polarity of word senses in WordNet produced via bootstrapping from a small manually compiled seed set. Each synset has two scores assigned, representing the positive, and negative polarity. The method used to develop Sentiwordnet is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification.

Table 2 shows the statistics of the adjective part of the English Sentiwordnet (SWN) in relation to the adjective part of the Dutch Wordnet (DWN). As can be seen from the first row, the English Sentiwordnet (18,563 synsets) is considerable larger than the Dutch wordnet (7,898 synsets). The wordnets are connected to each other by 17,754 translation equivalent links. Dutch translated synsets have an average of 3.1 translation links per synset.

Adjectives	**SWN**	**DWN**
number of synsets	18,563	7,898
translated synsets	8,217	5,725
equivalent links	17,754	17,754

Table 2. Statistics DWN and SWN

The transfer of the polarity values from the English to the Dutch wordnet consists of the following steps: (1) Copy the set of sentiment scores (positive and negative) from a SWN synset into the equivalent Dutch synsets (2) Calculate one set of scores for each DWN synset by counting up the positive scores and negative scores, respectively. As can be seen from table 2, many Dutch synsets have more than one translation equivalent which results in multiple sets of scores per synset. (3) Translate the two accumulated scores into one categorical value by attributing positive value if the positive_score is larger than the negative_score, and negative value if the positive_score is smaller than the negative_score. A synset is considered neutral if both scores are equal (being zero or larger than zero). (4) Assign neutral polarity to all synsets that are not covered by the transfer method , i.e. all synsets that do not have translation links with the English Wordnet.

To be able to compare the quality of the source and the target lexicons, Sentiwordnet1.0 was evaluated against the 'simplified' WNO-273 (cf. section 2.2). Different versions of Sentiwordnet have already been evaluated against Micro-WNOp by other studies (Baccianella et al. (2010)), but these evaluations use scalar values. For the present study, we converted the numerical scores of Sentiwordnet into categorical ones by applying the same rules as described above for the conversion of Micro-WNOp's numerical values.

3.2 Method I: Results and Discussion

The results of the transfer are presented in the following table. The first column (name) gives the name of the lexicon, e.g. SWN for the English Sentiwordnet and DSWN for the derived Dutch Sentiwordnet. The second column (gs) gives the gold standard against which the results are evaluated. The third column gives precision (P), recall(R) and weighted average (F) for all polarity (pol) categories together and for each one separately. By default, all other synsets are considered neutral and evaluated as such.

name	**gs**	**pol**	**P**	**R**	**F**
SWN (eng)	WNO-273	**All**	**0.62**	**0.62**	**0.62**
		POS	0.72	0.70	0.71
		NEG	0.58	0.63	0.60
		NTR	0.48	0.47	0.47
DSWN (dut)	ss-512	**All**	**0.58**	**0.58**	**0.58**
		POS	0.58	0.64	0.61
		NEG	0.61	0.61	0.61
		NTR	0.54	0.47	0.50
SWN-retro	WNO-273	**All**	**0.67**	**0.67**	**0.67**
		POS	0.74	0.85	0.79
		NEG	0.64	0.72	0.67
		NTR	0.54	0.31	0.40

Table 3. Evaluation Results English and Dutch Sentiwordnet

When comparing the scores of the source Sentiwordnet1.0 and the target Dutch resource, we see that overall performance drops with 4% precision (from 62% to 58%). Interestingly, however, precision scores of individual categories may also rise (cf. negative polarity which rises from 58% to 61%).

A closer look at the data shows that different factors affect the outcome. Conceivably, a substantial number of the errors may be due to incorrect annotations in the source lexicon. One single incorrect annotation in the source lexicon can affect large quantities of synsets in the target lexicon if they have many translation equivalent links. For example, more than ten Dutch synsets have a translation equivalent link with [comfy#a#1 comfortable#a#1] which is incorrectly tagged as 'negative'.

However, the transfer method has also positive

side-effects: if a word sense has many translation equivalents, incorrect annotations may be solved by correct ones. For example, *behulpzaam* (helpful) has 3 related English synsets which are correctly tagged 'positive' and one synset that is incorrectly tagged as 'neutral' (nice#a#7). The 'neutral' *nice* will in this case be overruled by the correct polarity values of the other synsets. The following experiment shows how powerful this multiple translation effect can be. We transferred the derived Dutch sentiwordnet back into English and replaced the scores of the translated English synsets (i.e. 8,217 synsets, cf. Table 2) with the newly obtained scores. Table 3 (SWN-retro) shows that both overall performance (from 0.62 to 0.67) and precision rates for each polarity category (from 0.72 to 0.74, from 0.58 to 0.64 and from 0.48 to 0.54 for positive, negative and neutral polarity, respectively) increase.

Finally, also the quality of the translation equivalent links show impact on the results. As the automatically generated translation equivalent links between the Dutch and the English Wordnet are not yet all manually corrected, the Dutch Wordnet consists of synsets with high quality – manually corrected – links and synsets automatically derived links. We divided the gold standard in synsets with manually correct links (202 items) and synsets with automatically derived links (303 items), and measured performance on the Dutch gold standard.We obtain 0.60 for the manually corrected items and 0.56 for automatically derived items which leads to the conclusion the quality of the derived Dutch sentiwordnet will increase when all translation links are manually corrected.

3.3 Method I: Conclusion

It seems that the transfer of coarse-grained sentiment like positive and negative polarity between wordnets of different languages can be done in a reliable manner, since the decrease in performance – after transfer - is rather low with 4%. Important factors that bear effect on the outcome are the quality of the source lexicon and the quality of the translation links.

Moreover, as demonstrated by translating the lexicons back and forth, the transfer process not only worsens but also improves the polarity scores.

4. Method II: Seed propagation

The seed propagation approach relies on the assumption that the concepts that are represented by synsets that are closely related by semantic links, have similar meaning and thus similar sentiment. Many versions of this approach have been implemented for English (Andreevskaia and Bergler (2006), Esuli and Sebastiani (2006)).

Also for Dutch a similar approach has been used by Jijkoun and Hofmann (2009). They generated, however, a word level polarity lexicon whereas our approach is aimed at generating a synset level lexicon.

4.1 Method II

We start with a set of seed synsets of known polarity (positive, negative and neutral) which is propagated through the wordnet making use of the lexical relations between synsets. The synset seed list is augmented during each iteration by adding near-synonym, antonym, hyponym and hypernym synsets. After each iteration the augmented list is used as seed list for the next step until convergence is achieved and no new synsets are added to the result list. The synsets that not have been added to the result list are considered 'neutral'. We did several experiments varying the type of lexical relations, the number of iterations, and the size and the composition of the seed list.

4.2 Seedlist Composition and Size

Andreevskaia and Bergler (2006) showed that the composition of the seed list has a considerable impact on the performance of the system. They did 58 runs of their sentiment tagging system on unique non-intersecting seed lists and found that the accuracy ranged from 47.6% to 87.5%. They attribute these variation to the fact that the used seed lists consisted of words, and not synsets or word senses, and that several words have both neutral and sentiment laden meanings whereas only one of them is included.

We think, however, that this is not the only reason for the variation, but that also the size and composition of the seed list are of considerable importance. To test this, we did experiments with three different seed lists: a high quality one, a low quality one and a large one of mixed quality.

- a 'high quality' seed synset list (sds-HQ)

Our hypothesis is that a carefully selected list of seed synsets taking into account the number of lexical relations (synonyms, near_synonyms, antonyms, hyponyms) with other members, may produce better results than a randomly chosen seed list. A large number of semantic ties with other members in the field proves that the involved synsets represent sentiment bearing concepts that are central and prototypical (Andreevskaia and Bergler (2006)). Thus, core members are identifiable in a wordnet by the number of lexical relations (LRs) links they have. This assumption is confirmed by the fact that typical evaluative sentiment bearing words have many synonym links as they tend to group together in large synsets, as shown by Maks and Vossen (2010a). A 'high quality' seed synset list is composed, as follows: (1) select 250 adjective synsets with more than 8 LRs (2) annotate this list manually with positive, negative and neutral polarity and (3) exclude synsets that have synonyms with mixed – positive, negative and/or neutral polarity – members as they produce noise because of their ambiguity.

- a 'low quality' seed synset list (sds-LQ)

A seed list of equal size but 'low' quality is composed. This list includes 250 synsets which have less than 3 LRs.

- a large seed synset list (sds-GI)

To complete the experiment we produced a large seed synset list of mixed quality. We use the General Inquirer Lexicon (Stone et al., 1966) as the starting point for this seed list. The list consists of 2,558 unique adjective words with neutral (1,203), negative (800) or neutral (771) polarity. We then use the online Google translation service to translate this list of words into Dutch. The seed words are related to the appropriate synsets. This procedure results in 1,411 labeled Dutch synsets, (428 neutral, 422 positive and 561 negative) 'of mixed quality'. The list includes both low quality seeds with less than 3 LRs (315 synsets) and high quality seeds with more 8 LRs (322 synsets).

seeds	ss-512	ss-complement	
sds-HQ	0.69	0.65	414 synsets
sds-LQ	0.55	0.55	498 synsets
sds-GI	0.75	0.67	236 synsets

Table 4 propagation with different seed lists

Table 4 shows the results obtained after propagation of the seed lists through the wordnet. The results have been evaluated against the complete gold standard (column ss-512) and against reduced versions of the gold standard from which the intersection between gold standard and seed list is removed resulting in 3 different test sets of 498, 414 and 236 items respectively (cf. Column ss-complement). By doing both evaluations we know if scores are due to larger overlaps of manually annotated seed list items and gold standard items or if they may be ascribed to the quality of the seed list.
The scores confirm our hypothesis that the number of LRs is indicative for the performance: the sds-HQ scores better than sds-LQ on both the full test set and the reduced version (0.69 vs. 0.65 and 0.65 vs. 0.55, respectively). However, the large seed list (sds-GI) performs even better and outperforms the high quality list on both versions of the test set. The fact that sds-GI scores better than sds-HQ even on the reduced version (0.67 vs. 0.65), suggests that the number of seeds might be even more important than the quality.
For further experiments with the propagation algorithm (cf. following sections) we use the sds-GI as it is the best scoring seed list.

4.3 Polarity Values

The performance of the propagation algorithm differs with regard to the different polarity categories (cf. Table 5.

Seeds	Gs	pol	P	R	F
sdsGI	ss-512	**All**	**0.75**	**0.75**	**0.75**
		POS	0.78	0.76	0.77
		NEG	0.76	0.82	0.79
		NTR	0.72	0.68	0.70

Table 5: Performance of different polarity categories

The scores of the neutral items, especially recall, are lower than those of the sentiment laden items. This is probably due to the fact that, although the number of seeds is almost equal for the different polarity categories, neutral items have less quality (cf. previous section) than the other categories as they have fewer lexical relations. Table 6 shows that 428 neutral seeds have an average of 2.5 synonyms (column SYN) and 2.5 other lexical relations (column SAHH: near-synonyms, antonyms, hypernyms and hyponyms) per synset whereas the negative and positive seeds have an average of 3.4 to 3.6 for both.

Pol	nr of seed synsets	SAHH	SYN
POS	422	3.4	3.4
NEG	560	3.6	3.5
NTR	428	2.5	2.5

Table 6 Average of LRs per synset

4.4 Number of Iterations

We experimented with the number of iterations (i) given in the first column of Table 7. Best balance between precision and recall is achieved with 5 iterations. With 10 iterations convergence is achieved. It is this last setting that is used throughout this paper.

i	Gs	Pol	P	R	F
0	-ss-512	**All**	**0.64**	**0.64**	**0.64**
		POS	0.87	0.51	0.64
		NEG	0.81	0.58	0.67
		NTR	0.49	0.85	0.62
1		**All**	**0.73**	**0.73**	**0.73**
		POS	0.81	0.70	0.75
		NEG	0.78	0.74	0.76
		NTR	0.63	0.74	0.68
5		**All**	**0.76**	**0.76**	**0.76**
		POS	0.78	0.74	0.76
		NEG	0.77	0.81	0.79
		NTR	0.71	0.70	0.71
10		**All**	**0.75**	**0.75**	**0.75**
		POS	0.78	0.76	0.77
		NEG	0.76	0.82	0.79
		NTR	0.72	0.68	0.70

Table 7: Various numbers of iterations (I)

With each iteration, recall increases while precision decreases as far as negative and positive polarity items are concerned. Varying the number of iterations can thus be used to produce small lists of lexical units with high

precision rates with regard to positive and negative sentiment.

4.5 Lexical Relations (LRs)

In order to propagate the seeds through the wordnet near_synonym (comparable to similar_to in Princeton Wordnet), antonym, hyponym, and hypernym relations are used. The adjective part of the Dutch Wordnet includes 3,119 hypernym/hyponym relations, 1,070 antonym relations and 703 near_synonym relations. The hierarchy is rather flat with many top nodes and only few synsets that have both hypernym and hyponym relations.

	Lexical relation	F
1	Ant(onym)	0.66
2	Hyper(nym)	0.66
3	Syn (near synonym)	0.67
4	Hypo(nym)	0.71
5	Syn+Ant+Hyper	0.69
6	Hyper+Hypo	0.73
7	Ant+Hypo+Hyper	0.74
8	Syn+Ant+Hypo	0.75
9	Syn+Hypo+Hyper	0.75
10	Syn+Ant+Hypo+Hyper	0.75

Table 8 Various types of Lexical Relations

Table 8 (row 4) shows that the best scoring relation is the hyponym relation with 0.71 whereas the other relations (cf. row 1-3) hardly outperform each other. Combinations of links score equally good (0.75) as long as the near-synonym (Syn) and hyponym (Hypo) relations are included (cf. row 8-10). When all relations are used, the impact of the hypernym relations is nihil (cf. row 8 and 10). The same holds for the antonym relation (cf. row 9 and 10): when all other relations are used the antonym relations do not affect the outcome.

Our conclusion is that there are no LRs which decrease performance. The combination of LRs scores best but only until a certain limit is reached.

These results will differ between wordnets. For example, as in the Princeton Wordnet there are no hyponym/hypernym relations between adjectives, the existing lexical relations will score differently.

4.6 Method II: Conclusions

We conclude that the performance of the propagation approach is determined by the number of iterations, the type and number of lexical relations and the type of seed list. The most important factor in determining the outcome of the propagation algorithm is the size of the seed list, i.e. the larger the better. Another important factor is the quality of the seed list; we proposed a set of rules which can be used to compile a well reasoned seed list.

5 Comparison of Method I and Method II

seed set	lexicon items	F
Transfer	Synsets	0.58
propagation-sdsGI	Synsets	0.75
propagation-sdsHQ	Synsets	0.65
propagation-sdsLQ	Synsets	0.55
Combi II + I	Synsets	0.74

Table 9 Results of transfer (I) and propagation (II) method

The results (copied in Table 9 from earlier sections for reader's convenience) show that the propagation method performs better than the transfer method. The results of the propagation method (0.75) outperform the transfer method (0.58) with 17%. Only with regard to the short and low quality seed list (0.55), the transfer method performs better than the propagation method.

We already mentioned that the two methods might complement each other as they cover different parts of the Dutch Wordnet (cf. Section 2.1). Therefore, the results of the two methods are combined, by taking the scores of the – best scoring – propagation method and replacing that part (2,109 synsets) that lacks lexical relations with the scores of the transfer method. The results show that the overall score degrades with 1%. This means that for those synsets which lack lexical relations, the default value 'neutral' performs better than the transfer method.

6 Comparison with other polarity lexicons

- vs. a word level lexicon for Dutch

To be able to compare the synset level results with other word level polarity lexicons, we generate a word level version of our lexicon. The results are evaluated against the 1,916 Dutch positive, negative and neutral words of the gold standard w-1916 (cf. section 2.2) and have a performance of 74%. This means that the extra step to bring the polarity values from synsets to words causes a small decline (1%) only.

seed set	language	lexicon items	gs	F
sds-GI	Dut	Words	w-1916	0.74
UvaLex	Dut	Words	w-1916	0.72

Table 10: Results at word level

The scores are compared with the scores of Jijkoun and Hofmann (2009)) who built a polarity lexicon (UvaLex) for Dutch at word level. Their approach is also wordnet based and makes use of lexical relations like synonyms and antonyms and of word-to-word links. Their results are comparable (table 10, row UvaLex) with ours (0.72 vs.

0.74 F-measure). Interestingly, an approach like the one of Jijkoun and Hofmann (2009) which is aimed at polarity annotation at word level, and therefore uses word-to-word relations to propagate the sentiment through the wordnet, does not perform better on word level than our system which is primarily meant for synset level annotation.

- vs. an English word level polarity lexicon

Secondly, the word level results are compared with an English polarity lexicon. Andreevskaia and Bergler (2006) whose annotations are at synset level and then aggregated to the word level evaluated their results against General Inquirer (Stone et al., 1966), and report an overall precision of 66.5%, for all 22,000 adjectives in the English Wordnet. For a smaller selection of 1,828 words with positive or negative polarity only, they report 83% precision. This is comparable with our scores; if we make smaller selections by applying fewer iterations and focus on positive and negative polarity only, we measure 82% precision for 2,530 adjective words. So, overall scores for the complete English wordnet are considerably lower than for the complete Dutch wordnet but with regard to smaller selections, Dutch and English perform equally good.

- vs. an English synset level polarity lexicon

The English Sentiwordnet1.0 (2006) is the only freely available polarity lexicon which covers all synsets of the Princeton Wordnet. A more recent version, Sentiwordnet3.0 which has higher scores than the previous versions, but is not publicly available (Baccianella et al. (2010)).
We measured on Sentiwordnet1.0 an overall performance of 62% (cf. section 3.2 above) which is considerably lower than our scores (0.75 and 0.69 for both seed lists, respectively). However, also in the case of Sentiwordnet1.0, smaller selections produce better results. For example, on a selection of 1648 high scoring positive and negative synsets, 84% precision is achieved.

A weakness of this study is that the results are not tested against one single gold standard. However, since we want to compare lexicons of different languages and different lexicon items (words vs. synsets), this is clearly impossible. We think that observed differences between English and Dutch are due to the considerable difference in size of the English wordnet and the Dutch wordnet (18.563 and 7.898 adjective synsets respectively) . The assumption is supported by the fact that small selections of high scoring items perform equally good across the two wordnets.

7 Conclusions

In this paper we described two approaches to generate synset level polarity lexicons for Dutch. The first approach builds a Dutch language polarity lexicon by translating the English Sentiwordnet into Dutch using translation equivalent links between the Dutch and the English Wordnet. The second approach generates a Dutch polarity lexicon at synset level propagating a seed list of known seeds through the wordnet using lexical relations.

It seems that the transfer of coarse-grained sentiment like positive and negative polarity between wordnets of different languages can be done in a reliable manner, since the decrease in performance – after transfer - is rather low with 4%. Important factors that bear effect on the outcome are the quality of the source lexicon and the quality of the translation links.

However, in the case of the Dutch Wordnet, we found that the propagation method considerably outperforms the transfer method. The best scoring seed list is a large seed list of 1,411 seed synsets, but a smaller 'a high quality' seed synset list, i.e. a list of synsets with many lexical relations, produces rather high scores as well.

Another objective of our study was to find out how methods designed for generating a synset level polarity lexicon perform at word level. Our conclusion is that the differences between the word level and synset level results are so small that they may be considered negligible.

8 Acknowledgements

This research has been carried out within the project From Text To Political Positions (http: //www2.let.vu.nl/oz/cltl/t2pp/). It is funded by the VUA Interfaculty Research Institute CAMeRA

9 References

Andreevskaia, Alina and Sabine Bergler (2006). Sentiment Tagging of Adjectives at the Meaning Level. In *LNAI 4013:. Advances in Artificial Intelligence. 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI-2006*, Springer-Verlag, Heidelberg and Berlin, Germany.

Baccianella, S. , Andea Esuli, F. Sebastiani (2010) Sentiwordnet 3.0 : an enhanced lexical resource for sentiment analysis and opinion mining . Lrec,Malta

Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). *Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming)*, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

Esuli, Andrea and Fabrizio Sebastiani. (2006). SentiWordNet: A Publicly Available Lexical Resource

for Opinion Mining. In *Proceedings of LREC-2006*, Genova, Italy.

Fellbaum, Chrisrtiane (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Gyamfi, Y., J. Wiebe, R.Mihalcea, C. Akkaya (2009). Integrating knowledge for subjectivity sense labeling. In *Proceedings of HLT-NAACL2009,* Boulder, Colorado.

Jijkoun, V. and K. Hofmann (2009) Generating a Non-English Subjectivity Lexicon: Relations That Matter. In *Proceedings of EACL-2009*, Athens, Greece.

Maks, I.and P. Vossen (2010a) Modeling Attitude, Polarity and Subjectivity in Wordnet. In *Proceedings of Fifth Global Wordnet Conference*, Mumbai, India.

Maks, I. and P. Vossen (2010b) . Annotation Scheme and Gold Standard for Dutch Subjective Adjectives. In *Proceedings of LREC-2010.* Valletta, Malta.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Su, F.and K. Markert (2008). Eliciting Subjectivity and Polarity Judgements on Word Senses. In *Proceedings of Coling-2008*, Manchester, UK.

Su, F; Markert, K. (2009). Subjectivity Recognition on Word Senses via Semi-supervised Mincuts. In: *Proceedings of NAACL-2009:* Boulder, Colorado.

Wiebe, Janyce and Rada Micalcea.(2006) . Word Sense and Subjectivity. In *Proceedings of ACL'06,* Sydney, Australia.

Vossen, P., I.Maks, R. Segers and H. van der Vliet (2008). Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In *Proceedings of LREC-2008*, Marrakech, Morocco

Towards the Automatic Merging of Language Resources

Silvia Necsulescu†, Núria Bel†, Muntsa Padró†, Montserrat Marimon*, Eva Revilla†

† IULA
Universitat Pompeu Fabra
Roc Boronat 138,
08018 Barcelona

* Universitat de Barcelona*
Gran Via de les Corts Catalanes, 585
08007 Barcelona

E-mail: {nuria.bel, muntsa.padro, silvia.necsulescu, eva.revilla}@upf.edu, montserrat.marimon@ub.edu

Abstract

Language Resources are a critical component for Natural Language Processing applications. Throughout the years many resources were manually created for the same task, but with different granularity and coverage of information. To create richer resources for a broad range of potential reuses, information from all resources has to be joined into one. The high cost of comparing and merging different resources by hand has been a bottleneck for merging existing resources. With the objective of reducing human intervention, we present a new method for automating merging of resources. We have addressed the merging of two verb subcategorization frame (SCF) lexica for Spanish. The results achieved, a new lexicon with enriched information and conflicting information signalled, reinforce our idea that this approach can be applied for other task of NLP.

1. Introduction

The production, updating, tuning and maintenance of Language Resources for Natural Language Processing is currently being considered as one of the most promising areas of advances for the full deployment of Language Technologies. The reason is that these resources that describe, in one way or another, information about the characteristics of a particular language are necessary for language technologies to work. For many technologies –Machine Translation, Parsing, Information Extraction, etc.– this particular information is stated in the form of a lexicon that registers how words are used and combined within that language. In other cases, the technology induces this information from a corpus of texts annotated with explicit information about these relations. Thus, the demand of both annotated corpora and lexica has augmented in the last years.

Although the re-use of existing resources such as WordNet (Fellbaum, 1998) in different applications has been a well known successful case, it is not very frequent. The different technology or application requirements, or even the ignorance about the existence of other resources, has provoked the proliferation of different, unrelated resources that, if merged, could constitute a richer repository of information augmenting the number of potential uses. This is especially important for under-resourced languages (perhaps for all but English), which normally suffer the lack of broad coverage resources. The research reported in this paper was done in the context of the creation of a gold-standard of subcategorization frames of Spanish verbs to be used in lexical acquisition (Korhonen, 2002). We wanted to merge two hand-written, large scale Spanish lexica to get a new richer and validated one. Because subcategorization frames contain rich and structured information, it was considered a good scenario for testing language resource merging methods.

Several attempts of resource merging have been addressed and reported in the literature. Hughes et al. (1995) report on merging corpora with more than one annotation scheme. Ide and Bunt (2010) also report on the use of a common layer based on a graph representation for the merging of different annotated corpora. Teufel (1995) and Chan & Wu (1999) were concerned with the merging of several source lexica for part-of-speech tagging. The merging of more complex lexica has been addressed by Crouch and King (2005) who produced a Unified Lexicon with lexical entries for verbs based on their syntactic subcategorization in combination with their meaning as described by WordNet, Cyc (Lenat, 1995) and VerbNet (Kipper et al., 2000).

In this context, proposals such as the Lexical Markup Framework, LMF (Francopoulo et al. 2008) become an attempt to standardize the format of computational lexica as a way to avoid the complexities of merging lexica with different structures.

In what follows, we will first introduce some background information about SCF lexica, and describe each resource involved in the experiment. We will also demonstrate an issue of encoding: how the same phenomena can be represented differently in each lexica, and introduce the structure of features that will be merged. In section 3, we will present the merging process, analyze the results obtained and introduce the need of adjusting results. Finally, in section 4, we will draw conclusions from our experiment, and advance future lines of research to further pursue the goal of reducing human intervention to only at the verification step.

2. Information encoded in SCF lexica

Subcategorization frames (SCF) are meant to make explicit the number and role of the complements that a predicate, most typically a verb, needs for forming a correct sentence and, more importantly, being correctly interpreted. Thus, the interpretation of sentence “John eats every morning” crucially depends on the knowledge that the verb “to eat” can be intransitive, that is, there is no

need to take a noun phrase as a complement. Note that the most usual case is that one lemma has more than one SCF, as is shown in Table 2. For every instance of one lemma in a text, the corresponding SCF should be chosen regarding its complements. As we have seen in the last example, the meaning of a sentence is strongly related to the complements of the verb. The decision on whether or not an element is a complement of a particular verb is made by a syntactic analysis which implies a parser. Parsers must be supplied with information to describe the syntactic behavior of each verb such as the number and characteristics of the complements that every verb takes, whether the occurrence of these complements is obligatory or not, and on how every particular complement contributes to the meaning of the whole sentence. Currently, both rule-based and statistical parsers benefit from this lexical information, first in the analysis step and the latter in the learning process (Jurafsky and Martin 2009 and Manning and Schütze 1999, for a discussion of the benefits of lexicalized statistical parsing). It is important to note that SCF phenomena differ substantially among language families. For instance, for Romance languages to encode how verbs behave with respect to cliticization phenomena, including "se" pronominalization is mandatory.

In the experiment we report here, we merged two subcategorization lexica developed for rule-based grammars; the Spanish working lexicon of the Incyta Machine Translation system (Alonso, 2005) and the Spanish working lexicon of the Spanish Resource Grammar, SRG, (Marimon, 2010) developed for LKB framework (Copestake, 2002). Note that different senses under the same lemma are not distinguished in these lexica, and thus, are not addressed in the research reported here. In the case of one lexicon enriched with different senses for one lemma, the merging mechanism would be the same. The difference would stay in the lexicon indexation. Instead of grouping the SCFs with respect to a lemma, they will be grouped under each pair's lemma-sense. Following is a brief description on how these two lexica encode SCF information.

2.1. The encoding of SCF in the Incyta lexicon

In the Incyta lexicon, each verb entry is represented as a list of tags. The subcategorization information for each verb is encoded in the 'ARGS' feature as a parenthesized list of all the possible subcategorization patterns that a given verb can have, even if the different patterns imply a change in the meaning of the verb.

The information contained in the SCF includes a list of the possible complements, indicating for each of them the grammatical function ($SUBJ, $DOBJ, $IOBJ, $POBJ, $SCOMP, $OCOMP, $ADV), the phrase type that can fulfill each grammatical function ('N1' for noun phrase, 'N0' for clausal phrase,, 'ADJ' for adjective phrase) and the preposition required in the case of prepositional objects ($POBJ). In the case of clausal complements, the information is further specified, indicating the type of clause (finite, 'FCP', or non-finite, 'ICP') in the interrogative ('INT') or non-interrogative ('0') forms, and the mode ('SUB' or 'IND' in the case of a finite clause) or the control structure ('PIV $SUBJ', 'PIV $DOBJ', etc.), in the case of non-finite clauses. Incyta further specifies if one of the complements can be fulfilled by a reflexive and/or reciprocal pronoun ('$DOBJ APT RFX'). Apart from the number and type of the complements, the subcategorization pattern includes further characteristics, represented by the GFT tag (General Frame Test). For example, whether the verb is impersonal for weather-like verbs (LEX-IMPS T) or if it can take the "se" clitic (RFX), that is, pronominal verbs as explained in 2.3, or if it can occur in the form of an absolute past participle construction.

2.2. The encoding of SCF in the SRG lexicon

The SRG is grounded in the theoretical framework of Head-driven Phrase Structure Grammar, HPSG, (Pollard and Sag, 1994), a constraint-based, lexica-list approach to grammatical theory where all linguistic objects (i.e. words and phrases) are represented as typed feature structures. In the SRG lexicon, each lexical entry consists of a unique identifier and lexical type (one among about 500 types, defined by a multiple inheritance type hierarchy).

Verbs are encoded by assigning a type and adding specific information to the lexical entries. Verbal types are first distinguished by the value for the SUBJ-list. Thus, we have subtypes for impersonal verbs taking an empty SUBJ-list, verbs taking a verbal subject and verbs taking a nominal subject.

The feature COMPS is a list of the complements which specifies the phrase structure type of each complement; i.e. noun phase (NP), clause phrase (CP), prepositional phrase (PP), adjectival phrase (AP), adverbial phrase (ADV), and subject complement (SCOMP). Verbal complements are specified for their form (finite or infinitive), mode (indicative or subjunctive), and control or raising relation of verbal complements. Marking prepositions for some verbs are given in the lexicon itself, while for the others just the preposition's type is specified. Alternations of complements, as well as other valence changing processes that verb frames may undergo, are dealt with by the grammar rules, which are triggered by lexical feature-value attributes that encode whether a verb is, for instance, reflexive or pronominal.

2.3. Issues in information merging

It is evident from previous section, that the SRG and the Incyta lexica encode the same phenomena but in a slightly different way. For a task like automatic merging, information about the same facts must be represented exactly in the same way as to compare and decide whether (Crouch and King, 2005):

- it is the same information
- it is different information that should be kept in the resulting lexicon
- it is different information that points at some gap or inconsistency in one of the lexica, if not directly to an error

In addition to mere formal differences, i.e. different tags, there can be differences in the semantics of a given tag, i.e. one tag covers what in another dictionary covers two tags. One of the most complex cases we found was the

encoding of reflexive and pronominal verbs in both lexica. Now, we will briefly review the implications of this phenomenon, the complexity of representing it and how these two lexica encode it differently, which was one of the more interesting issues to study in the results of the merging experiments.

In Spanish (but also in other languages like French, Italian, Dutch, German, etc.) the presence of the reflexive pronoun triggers, in combination with different verbs, different interpretations related to diathesis and the number and interpretation of the arguments. The so called pronominal verbs are those that are lexically marked, which in some constructions occur with a pronominal clitic particle 'se', without referential value. The lack of referential value distinguish these constructions from other clitic occurrences like the expletive use of clitics to refer to an obligatory, but not mentioned, object such as (1), or a true reflexive occurrence like in (2).

1. *La* vi
   'I saw her'
2. *Me* lavo las manos
   'I wash my hands'

Pronominal verbs are normally classified into two groups. *Inherent/absolute pronominal verbs*: Their frame obligatorily requires the occurrence of the clitic and because it depends only on the lexical item, it must be encoded in the lexica.

3. Juan se ha atrevido a pedir un aumento
   'John CLI dared to ask for a raise'
   *Juan ha atrevido a pedir un aumento

Argument reducing pronominals: When appearing with the clitic, its otherwise transitive structure is reduced in one element and the internal argument becomes external. It is normally related to anticausativization phenomena (Bosque, 1999):

4. El capitán ha hundido su barco
   'The captain sank his ship'
5. El barco se ha hundido
   'The ship has sank'
6. Juan ha roto un vaso
   'John broke the glass'
7. El vaso se ha roto
   'The glass CLI broke'

In Spanish a further problem arises because of the surface similarity between these 'pronominal verb' constructions and the impersonal and reflexive passive sentences also expressed with the clitic 'se'. Most verbs can enter in these constructions where the 'impersonal' value comes from the fact that when appearing with 'se' they inflect in the 3rd. person, here is no lexical subject and they have not or they do not imply reference to any definite subject as they would do if the particle 'se' was eliminated.

8. Se vive bien en Barcelona
   'People live well in Barcelona'
9. Se han suspendido las negociaciones
   'The negotiations have been suspended'

In a reflexive passive construction the verb agrees in number with the nominal element which is considered grammatically to be the subject, producing thus a reduction in the number of complements too, like the pronominal case just mentioned.

Due to this variety of possible uses of "se" and the subtle nuances of their interpretations, there is a significant degree of hesitation, if not confusion, when encoding reflexive and pronominal verbs in the lexicon. Our two lexica were not an exception and, most probably because of the difficulties of consistently encoding pronominal verbs, each lexicon has opted for a different strategy and, critically, they do not always agree in the classification of a verb as pronominal or reflexive, the two cases where specific information in the SCF lexicon is required. The Incyta lexicon encodes the possibility of bearing a "se" clitic and taking part in an argument reduction phenomenon with the tag "GFT RFX" annotating the whole SCF. Besides, it marks the possibility of an argument taking a reflexive pronoun adding the feature-value "(APT RFX)" as an annotation in $DOBJ and $IOBJ complements.

The SRG lexicon distinguishes with different types among the reflexive or pronominal interpretation of a verb when occurring with "se".

	Reflexive			**Pronominal**		
	#verbs	#both	#singles	#verbs	# both	#singles
SRG	835	190	645	712	597	115
Incyta	204		14	1204		607

Table 1: Differences of reflexive and pronominal encoding in the two lexica

In table 1 we can see the number of verbs encoded as reflexive (as 1) and pronominal (as 4 and 5) and the overlapping of the two lexica expressed as the number of verbs equally encoded in both lexica. Singles refer to those that are only encoded as reflexive or pronominal in one of the lexica. Despite the difference in quantities, one can observe that the overlap is far from being in the majority, and that there is a significant amount of systematic differences within the encoding.

2.4. The encoding of SCF in the common lexicon

As we said before, our objective was to merge two SCF lexica by graph unification which allows us to combine the information contained in two lexica. This method fulfills our objective to create a complete and correct SCF lexica using information from two manually created resources. By unification, we validate the common information, exclude the inconsistent and add the unique information that each lexicon contains.

The first step of the process is to convert each lexicon into a format which supports graph unification. We decided to use feature-value structures, which form directed acyclic graphs, i.e. the features are arrows and the values, nodes. A graph being a structured representation, intuitively presents the lexical information and it can be easily

transformed, after merging, to other standard formats for further reuse.

The exercise of converting the information contained in a lexicon is referred to as the extraction phase and several rules were manually written according to the intended interpretation of the encoding found in the lexica in order to make it match only within the cases wanted, respecting different information that must occur in the new resource, and indicating when contradictory information occurs for the same verb.
The extraction phase revealed major differences between the two lexica in the following cases:

(i) Different information granularity. This was the case of the Incyta tag "N0" for referring to the category of the phrase that can fulfill a complement. It had to be split according to their form, into a 'finite' or 'infinitive' clause in order to compare with the SRG encoding.

(ii) Different grammatical coverage. For instance, the Incyta lexicon lists bound prepositions, while the SRG lexicon can refer to the type of the bound prepositions (i.e. locative or manner).

(iii) Different treatment of systematic complement alternations. SRG handled them by lexical rules while Incyta explicitly declared them as possible SCF or disjunctions included in one of them. For example, a verb that has a complement that may be fulfilled by both a finite and an infinitive clause is represented with a type that includes a lexical rule that will produce the alternation when needed. In the Incyta lexicon this phenomenon is encoded as two different realizations in the SCFs, one for the finite clause (FCP) and one for the infinite (ICP). Thus, in this example, one extraction rule would convert one SRG frame into two: one with finite and one with an infinite clause complement.

These differences in encoding resulted in a different number of SCF, which we will comment upon later.

In general terms, the extraction rules mapped the information of each lexicon into a graph that can be represented as an attribute-value matrix. The attribute and values used are the following (the names are used for internal purposes, but a translation into recommended LMF labels is planned):

- 'subj' specifies the category of the subject, i.e. Noun Phrase (NP), Complementizer Clause Phrase (CP);
- 'comp_1' and 'comp_2' specify the category of the first, respective to the second, verb complement, i.e. none (no complement), adverbial (adv), indirect object (ppa), adjectival (adj), NP, CP or PP.
- 'passive' specifies if the verb accepts to undergo passive.
- 'apc' specifies if the verb occurs as absolute past participle construction
- 'rpc' specifies if the verb is reciprocal verb as in the example: "Juan y María se escriben cartas" ('*Juan and María write letters to each other*').
- 'rfx_prn' is a complex valued attribute that specifies if the verb is reflexive ([clitic='yes'; rfx='yes'; prn='no']), pronominal ([clitic='yes'; rfx='no'; prn='yes']) or none of them ([clitic='no'; rfx='no'; prn='no']).

The attributes 'subj', 'comp1' and 'comp2' can be simple structures, i.e. NP, or complex structures. In the latter case, they include a list of specific features, indicating the category, their form (finite or infinite, affirmative or interrogative), the verbal mode (indicative or subjunctive) or the preposition required.

Another complex structure is the 'rfx_prn' attribute containing the 'clitic' feature discussed in section 2.3, which takes 'yes' if the verb is a pronominal or reflexive and 'no' if it does not accept this type of pronoun. 'prn' and 'rfx' information is triggered by the SRG lexicon which encodes, whether a verb is pronominal or reflexive. For verbs which accept both types of pronouns, they will have two different SCF, one for each behavior.

As we see in table 1, the agreement among the two lexica regarding the encoding of reflexive and pronominal phenomena was far from being complete. This would be a handicap for unification algorithms which unify only if the values of all common features are compatible (in this case if the verb is reflexive or pronominal). Thus, we had to standardize the division in reflexive and pronominal classes from one of the lexica. After a manual inspection we decided to preserve the SRG information because it was richer, we collapsed the Incyta information in only one feature "clitic=yes", i.e. the verb is pronominal or reflexive, and let the SRG make the final decision during the moment of unification. For each verb in the SRG, a further "clitic=yes" was added, at the same level as 'prn' and 'rfx' features, to prevent unification with those entries that had no information, and thus can unify without restrictions.

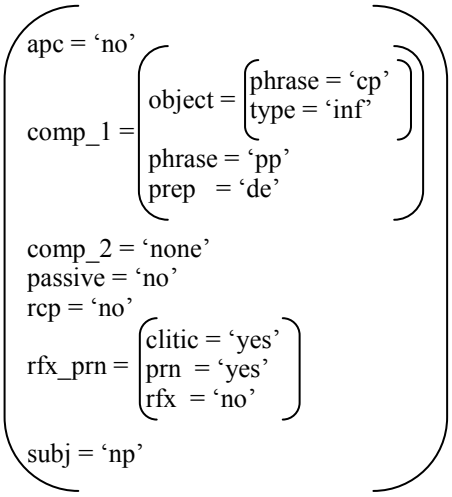
When we organize information from both dictionaries in a common format, we looked for a structure that keeps the information valid during the process of graph unification. For instance, a PP with a CP object cannot unify with a PP with a NP object.

3. Unification

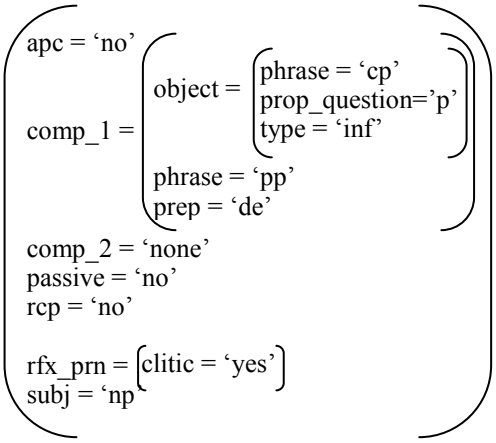
After the manual effort of conversion into a ready to unify format, the second step was the unification of the two lexica that contain the same structure and features. The objective of merging the two SCF lexica was to have a new, richer lexicon with information coming from both. After each lexicon was mapped into a common format, the results were mechanically compared and combined to form the new resource.

Once the SCF was converted into graphs, we used the basic unification mechanism implemented in NLTK (Bird et al., 2009) for each verb to merge its SCF from the Incyta lexicon with those from the SRG lexicon. For a better understanding of the unification process, in Figure 1 we present the results of the unification for the verb 'reprimir' (to repress), where it is interesting to note the resulting values of the 'rfx_prn' feature. This verb is considered in the Incyta lexicon as a 'clitic' verb, without

SRG:



Incyta:



Result:

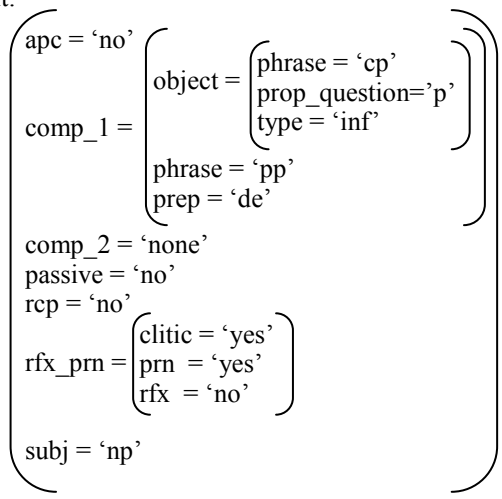


Figure 1: The results of the merging for the verb 'reprimir' (to repress).

expressing the values for 'rfx' and 'prn' features. However, in the SRG lexicon, it is encoded as a pronominal verb; therefore the final SCF lexicon considers it also a pronominal verb. In addition, the same example presents a case of lack of information in SRG because it does not specify the differences of the causal phrase from the PP in the case that it is a statement or a question. The resulting structure fulfills our objective to maintain information from both lexica.

The unification process tries to match many-to-many SCFs under the same lemma. This means that for every verb, each SCF from one lexicon tries to unify with each SCF from the other lexicon. The resulting lexicon is richer in SCFs for each lemma, on average, as shown in Table 2, where we present the results of merging the two lexica in terms of SCFs, lemmas and the average of SCF per lemma. Note that we present both the number of unique SCFs in the three lexica and the number of total SCFs that can be found in them.

The resulting lexicon will contain lemmas from both dictionaries and for each lemma, the unification of the SCFs from the Incyta lexicon with those from the SRG lexicon. The unified SCFs can be split in three classes:

(1) SCFs of verbs that were present in both dictionaries, i.e. A_{SCF} is contained under one lemma in both lexica, thus the resulting lexicon, contains A_{SCF} under this lemma;
(2) Information on SCF's components that were present in one of the lexicon but not in the other, i.e. the Incyta lexicon contains $A_{SCF,}$ while the SRG lexicon contains B_{SCF} under the same lemma. A_{SCF} and B_{SCF} unify in C_{SCF} , where C_{SCF} contains the common information and also the information just in A_{SCF} or just in B_{SCF}.;
(3) SCFs that were present in one of the lexicon but not in the other: the Incyta lexicon contains A_{SCF}, while the SRG lexicon contains B_{SCF} under the same lemma. A_{SCF} and B_{SCF} cannot unify, thus the resulting lexicon contains for the same lemma both frames, A_{SCF} and B_{SCF}.

Group (3) consists in inconsistent information in lexica, as it can signal a lack of information in one lexicon (e.g. A_{SCF} appears in Incyta but it does not have a corresponding SCF in SRG) or an error in the lexica (at least one of SCF implicated into the unification is an incorrect frame for its lemma). Thus, for detection conflicting information, we will detect lemmas whose SCFs do not unify at all (the unification number under a lemma is 0), or SCFs in one or the other lexicon that never unify with other SCFs (the total unification number for a SCF is 0). In a further step, by using a human specialist, this information can be manually analyzed and eventually eliminated from the final lexicon. Our objective is to automatically merge a lexica, thus we consider human analysis a possible intervention that would be useful to filter the results, but not a necessary step. The resulted lexicon contains all valid information provided by the unification of lexica and some SCFs that can be incorrect or not.

Lexicon	Unique SCF	Total SCF	Lemmas	Avg.
SRG	326	13.864	4303	3.2
Incyta	660	10.422	4070	2.5
Merged	919	17.376	4324	4

Table 2: Results of the merging exercise

It can be seen from the number of unique SCFs that the Incyta lexicon has many more SCFs than the SRG lexicon. This is due to different granularity of information. For example, the Incyta lexicon always gives information about the concrete preposition accompanying a PP while, in some cases, the SRG gives only the type of preposition, as explained before. The number of unique SCFs of the resulting lexicon, which is close to the sum between the numbers of the unique SCFs in the lexica, was very surprising for us. As shown in Table 3, for 50% of the lemmas we have a complete unification; thus this result comes from the many to many unification rather than from the direct addition of SCFs from both lexica.

Regarding the average number of SCFs per lemma in the different lexica, we use the total number of SCFs to calculate it.

Lemmas	Unification classes (lemmas)	Resulted SCF				
		Total	Unify	No unify		
				SRG	Incyta	
4050	2166	3329	3329	0	0	(1)
	888	7424	1966	3119	2339	(2)
	525	2977	1123	1854	0	(3)
	197	991	600	0	391	(4)
	274	1810	0	1123	687	(5)
274		845	0	778	67	(6)

Table 3: Detailed results of merging: (1) Unify 100%; (2) There are not unified SCFs in both lexica; (3) There are not unified SCFs in SRG; (4) There are not unified SCFs in the Incyta lexicon; (5) Any SCFs do not unify; (6) Appears only in one lexicon.

Table 3 explains with more details the source of this gain of SCFs. Our final lexicon contains a total of 4,324 lemmas. From those, 4,050 appeared in both lexica (94%). 2,166 lemmas (class (1) from Table 3) unified all their SCFs signifying a total accord between both lexica for 50% of lemmas. Note that for 2,160 of them, every SCF from the Incyta lexicon unifies with one and only one SCF in the SRG lexicon, that is a unification type '1 to 1', while 6 verbs accomplish a many-to-many unification.

1,610 (the classes (2), (3) and (4) from Table 3) lemmas do not unify all the SCFs thus they reveal differences between both lexica, as explained in section 2.4. These lemmas present, in total, 8,637 SCFs in the SRG lexicon and 6,342 SCF in the Incyta lexicon. Through the unification process under the same lemma, 3,689 SCFs unify, while a total of 4,973 SCFs from the SRG lexicon and 2,730 SCFs from the Incyta lexicon are added directly into the resulting lexica. Besides, the resulting lexicon contains 274 lemmas (the class (6) from Table 3) that appear just in one lexicon, 21 lemmas appear just in the Incyta lexicon and 253 lemmas appear just in the SRG lexicon, which are considered as lacking of information. They are the best proof of our results that the new lexicon is more consistent in information.

Only 274 lemmas, 6,3%, did not unify any SCFs because of conflicting information and require further manual analysis. An example of complete unification failure comes from the inconsistent encoding of pronominal and reflexive verbs in a hand-made lexicon like the one we have introduced in section 2.3.

In order to assess the quality of the new resource, we performed a manual inspection of lemmas whose frames can't be unified. Our objective was to identify what was the inconsistent information and we had a special interest in the results of the merging of the pronominal and reflexive verbs, which we knew are problematic.

In the Incyta lexicon, most of the reflexive or pronominal verbs have two different SCFs: one for the occurrence of the clitic personal pronoun, no NP complement and the tag for a reflexive verb (e.g. cubrir: "yo me cubro", *I cover myself*) and another one for the NP complement, in this case it is no longer encoded with the tag for reflexive verb ("yo cubro el coche", *I cover the car*).

On the contrary, in SRG, both realizations of a reflexive verb are included in the same frame, indicating both that it may have a NP complement and a reflexive tag. Because the clitic pronoun and the NP complement cannot appear in the same SCFs, when extracting all possible SCFs that a SRG verb may have our set of extraction rules creates two SCFs: one reflexive, without NP complement, and another non-reflexive with NP complement. Using this strategy, we obtained over 3600 unifications for these types of verbs, thus we consider our approach correct. However, we found that around 100 Incyta verbs had been encoded following the same interpretation as the original the SRG lexicon. These verbs have a SCF that contains both the NP complement and the reflexive tag and thus do not unify with the SRG SCF's that have been split into two SCFs. These verbs are a third of the ones that do not unify any SCF. Figure 2 demonstrates these particular feature structures.

As it can be seen from the tables above, the resulting lexicon is richer than the two it is composed of as it has gained information in the number of SCFs per lemma, as well as in the information contained in each SCF. Table 2 shows an increase of SCFs per lemma on average.

In general, automatic merging produces errors that can easily be the object of further refinement, because errors are systematic. However, this tends not to be true of manual merging exercises, where human errors are occasional and hence, inconsistent, as we have seen in the encoding of reflexive verbs in the Incyta lexicon.

Thus, an automatic merging process can have a final step, based on what Crouch and King (2005) call "patch files". Using our observations collected during the final verification, we will consider for the future to devise specific patches that correct or add information in particular cases where either wrong or incomplete information is produced. A first candidate case would be to correct all of the verbs in the Incyta lexicon with SCFs that have both the reflexive tag and the NP complement.

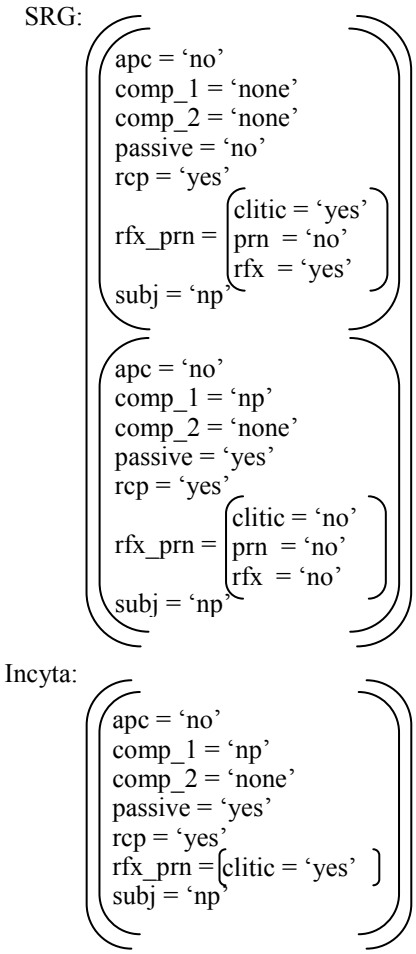


Figure 2: Example of unification problem for a reflexive verbs, such as 'cubrir' ('to cover')

4. Conclusions

We have proposed a method to reduce human intervention in the merging of Language Resources, in particular within the SCF lexica. By using graph unification as the sole operation that controls merging, we support the proposal of Ide and Bunt (2010) for rich annotated corpus merging, demonstrating that it is also possible for lexical merging. Our proposal of extracting information and representing it as a graph in order to only use a unification method for the actual merging is an innovative proposal in the field of dictionary merging. The structure proposed is based on attribute-value feature-based directed acyclic graphs and can be easily transformed into a standard format for further reuse.

We consider the results obtained in our experiments very satisfactory. Unifying two SCF lexica after converting them to a common operative format by using extraction rules led to a richer resource that will be offered to the community as a gold-standard of verbal SCF for Spanish. During the unification step errors, which are systematic due to the formal merging process, can be detected and then corrected using patch files.

The mapping of information to a common structure remains a very expensive part of resource merging if done manually. It is future work to reduce the cost of information comparison and extraction exercises by proposing an automatic mapping solution.

5. Acknowledgments

This project has been funded by the PANACEA project (EU-7FP-ITC-248064) and the CLARA project (EU-7FP-ITN-238405).

6. References

Alonso J.A., Bocsák A. (2005). Machine Translation for Catalan-Spanish. The Real Case for Productive MT; In *Proceedings of the tenth Conference on European Association of Machine Translation (EAMT 2005)*, Budapest, Hungary.

Bird S., Klein E., Loper E. (2009) *Natural Language Processing with Python*. O'Reilly Media, 1 edition

Bosque I, Demonte V., Eds. (1999): Gramática descriptiva de la lengua española, R.A.E. - Espasa Calpe, Madrid.

Chan D. K., Wu.D. (1999). Automatically Merging Lexicons that have Incompatible Part-of-Speech Categories. *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*. Maryland.

Copestake A. (2002). Implementing Typed Feature Structure Grammars. *CSLI Publications*, CSLI lecture notes, number 110, Chicago.

Crouch D, King T.. (2005). Unifying lexical resources. Proceedings of *Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbruecken; Germany.

Farrar S, Langendoen D. T. (2003) A linguistic ontology for the Semantic Web. *GLOT International.* 7 (3), pp.97-100

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

Francopoulo G., Bel N., George M., Calzolari N, Pet M., Soria C. (2008). Multilingual resources for NLP in the lexical markup framework (LMF). *Journal of Language Resources and Evaluation*, 43 (1).

Hughes J., Souter C., Atwell E. (1995). Automatic Extraction of Tagset Mappings from Parallel-Annotated Corpora. *Computation and Language.*

Ide N. and Bunt H.. (2010). Anatomy of Annotation Schemes: Mapping to GrAF. *Proceedings of the Fourth Linguistic Annotation Workshop*, ACL 2010

Jurafsky D., Martin J.H. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, *Speech Recognition, and Computational Linguistics. 2nd edition*. Prentice-Hall.

Kipper K., Hoa Trang Dang, H.T., Palmer M.. (2000). Class-based construction of a verb lexicon. In *Proceedings of AAAI/IAAI.*

Korhonen A. (2002). Subcategorization Acquisition. PhD thesis published as *Technical Report UCAM-CL-TR-530*. Computer Laboratory, University of Cambridge

Lenat D. (1995). Cyc: a large-scale investment in knowledge infrastructure. *In CACM 38, n.11*.

Manning C.D., Schütze H. (1999). Foundations of Statistical Natural Language Processing. *MIT Press*,

Cambridge, MA, USA.

Marimon M. (2010). The Spanish Resource Grammar. Proceedings of *the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Paris, France: European Language Resources Association (ELRA).

Molinero Miguel A., Sagot Benoît and Nicolas Lionel 2009). Building a morphological and syntactic lexicon by merging various linguistic resources. In Proc. of 17th Nordic Conference on Computational Linguistics (NODALIDA-09), Odense, Danemark.

Pollard, Sag I.A. (1994). Head-driven PhraseStructure Grammar. The University of Chicago Press, Chicago.

Teufel S. (1995). A Support Tool for Tagset Mapping. *In EACL-Sigdat* 95

A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers

Alexis Amid Neme

Laboratoire d'informatique Gaspard-Monge – LIGM
Université Paris-Est, 77454 Marne-la-Vallée Cedex 2, France.

http://infolingu.univ-mlv.fr

E-mail: alexis.neme@gmail.com

Abstract

We describe a lexicon of Arabic verbs constructed on the basis of Semitic patterns and used in a resource-based method of morphological annotation of written Arabic text. The annotated output is a graph of morphemes with accurate linguistic information. An enhanced FST implementation for Semitic languages was created. This system is adapted also for generating inflected forms. The language resources can be easily updated. The lexicon is constituted of 15 400 verbal entries.

We propose an inflectional taxonomy that increases the lexicon readability and maintainability for Arabic speakers and linguists. Traditional grammar defines inflectional verbal classes by using verbal pattern-classes and root-classes, related to the nature of each of the triliteral root-consonants. Verbal pattern-classes are clearly defined but root-classes are complex. In our taxonomy, traditional pattern-classes are reused and root-classes are simply redefined.

Our taxonomy provides a straightforward encoding scheme for inflectional variations and orthographic adjustments due to assimilation and agglutination. We have tested and evaluated our resource against 10 000 diacriticized verb occurrences in the Nemlar corpus and compared it to Buckwalter resources. The lexical coverage is 99.9 % and a laptop needs two minutes in order to generate and compress the inflected lexicon of 2.5 million forms into 4 Megabytes.

1. Introduction

Arabic morphology can be described by many formal representations. However, Semitic morphology or *root-and-pattern* morphology (Kiraz, 2004) is a natural representation for Arabic [1]. The *root* represents a morphemic abstraction, usually for a verb a sequence of three consonants, like *ktb*. A *pattern is* a template of characters surrounding the root consonants, and in which the slots for the root consonants are shown by indices. The combination of a root with a pattern produces a surface form. For example, *kataba* and *yakotubu* are represented by the root *ktb* and the patterns *1a2a3a* or *ya1o2u3u*.

Root-and-pattern morphology is standard in Arabic and is learned in grammar text books. Arabic linguists use *root-and-pattern* representation in order to list verbal entries and related inflected forms. On the other hand, FSTs have shown their simplicity and efficiency in inflectional morphology for western languages. Computer scientists appoint FSTs as standard devices for inflection. Various formal representations for Arabic morphology have been created by computer scientists to avoid root-and-pattern representation. The point that motivated this trend is that FSTs formalism would not be fitted for Semitic morphology since FSTs are concatenative whereas Semitic morphology is not. In concatenative representation, the root-and-pattern representation is replaced by a stem- or lexeme-based representation. For these formalisms, a stem is a basic morpheme that undergoes affixations with other morphemes in order to form larger morphological or syntactic units. For root-and pattern morphology, a stem derives from a root and a particular pattern and subsequently undergoes affixations.

At the operational level, the lexical representation of the concatenative model is entirely concatenative in order to compel with the *[prefix][stem][suffix]* representation. However, these representations imply a manual stem precompilation based on a root-and-pattern representation. The concatenative models are generally composed of three components: lexicon, rewrite rules, and morphotactics. The lexicon consists of multiple sublexica, generally *prefix*, *stem*, and *suffix*. The rewrite rules map the multiple lexical representations to a surface representation. The morphotactics component aims with a subjacent representation to generate or to parse the surface form *[prefix][stem][suffix]* and performs alternation rules at morpheme boundaries such as deletion, epenthesis, and assimilation.

Any formal representation that is not adapted to Semitic morphology will be rejected by the majority of Arabic-speaking linguists. When linguists work in a newly created formalism, they continue to work with *root-and-pattern* representation on paper and subsequently, they unfold their descriptions for a specific formalism. Their contribution for updating and correcting lexical resources is complex and time-consuming, and therefore error-prone.

Our approach resorts to classical techniques of lexicon compression and lookup in an inflected full-form dictionary that includes orthographic variations related to morpheme agglutination. The formalization of all possible verbal tokens requires complex and interdependent rules. For these issues, we define a taxonomy for Arabic verbs composed of 460 inflectional

1 We would like to thank Eric Laporte and Sébastien Paumier for helpful discussions, contributions and for the adaptation of Unitex to Arabic. Unitex is an open source multilingual corpus processor. More than 12 European languages, Korean and Thai with their linguistic resources are operational in Unitex. http://www-igm.univ-mlv.fr/~unitex

classes. We demonstrate that FSTs are compatible with root-and-pattern representation. Our taxonomy encodes simultaneously in the lexical representation three variations at the surface level:
- inflectional classes of a lemma;
- inflectional subclasses related to morphophonemic assimilation;
- orthographic adjustments related to the agglutination of a pronoun.

In our orthographic representation, we use a fully diacriticized lexicon and we take advantage of the clear boundary, already defined in traditional grammar, between verbal inflection and verbal agglutination to describe these two levels independently. In order to satisfy both computer scientists and Arabic linguists, we have created in Unitex an enhanced version of FSTs adapted to root-and-pattern representation.

In Section 2, we outline the state-of-the-art approaches to Arabic morphological annotation. Section 3 describes the methodology and particularly the inflectional verbal taxonomy. Section 4 describes agglutination as morpheme combinatorics. Section 5 reports the construction of the lexicon. Section 6 reports the evaluation of the lexicon. A conclusion and perspectives are presented in Section 7.

2. State of the Art

Several morphological annotators of Arabic are available. The Buckwalter Arabic Morphological Analyzer (BAMA) is one of the best Arabic morphological analysers and is available as open source. The BAMA uses a concatenative lexicon-driven approach where morphotactics and orthographic adjustment rules are partially applied into the lexicon itself instead of being specified in terms of general rules that interact to realize the output (Buckwalter, 2002).

The BAMA has three components: the lexicon subdivided in A, B, C sublexica, the compatibility tables (AB, BC, AC) and the analysis engine. An Arabic word is viewed as a concatenation of three regions, a prefix region (A), a stem region (B) and a suffix region (C). The prefix and suffix regions can be null. An entry in A may be the concatenation of proclitics and an inflectional prefix. An entry in C may be the concatenation of an inflectional suffix and an enclitic. The A and C lexica are composed of 561 and 989 entries which represent all possible combinations of inflectional and agglutinative morphemes for nouns and verbs. For each stem in B, a morphological compatibility category, an English gloss and part-of-speech (POS) data are specified. A list of stems is assigned to a lemma, and the lemma is not used in the analysis process. The B lexicon is composed of 82 000 stems which represent nearly 40 000 lemmas. Verbal stems are 33393[2] and represent 8709 verbal lemmas. A full ABC form must be allowed by the three compatibility tables AB, BC, AC.

2 Verbal stems are for perfect active (17008) stems, imperfect active (13241), perfect passive (403), imperfect passive (2611), and for imperative 130 stems. BAMA resource does not include all imperfect active stems, for instance.

qr>	qara>	PV->	qara>/VERB_PERFECT
qr|	qara|	PV-|	qara|/VERB_PERFECT
qr&	qara&	PV_w	qara&/VERB_PERFECT
qr>	qora>	IV	qora>/VERB_IMPERFECT
qr>	qora>	IV_wn	qora>/VERB_IMPERFECT
qr|	qora|	IV-|	qora|/VERB_IMPERFECT
qr&	qora&	IV_wn	qora&/VERB_IMPERFECT
qr}	qora}	IV_yn	qora}/VERB_IMPERFECT
qr>	qora>	IV_Pass	yuqora>/VERB_IMPERFECT

Table 1. BAMA stem lexicon using Buckwalter transliteration. A list of stems related to the lemma-identifier qara>-a_1 "to read". The 9 stems are related to the orthographic variants of the 3rd root consonant, here glottal stop (*hamza*), depending on the next inflectional suffix and the existence of an agglutinated pronoun.

The Buckwalter representation for the Arabic lexicon is not fitted for generation but only for text analysis. In ElixirFM (http://elixir-fm.sourceforge.net/), Smrz (2007) adapted the Buckwalter resources for generation and the project is implemented in Haskell, a functional programming language. In the ALMORGEANA project, Habash (2004) proposed also a version of Buckwalter resources adapted to generation and analysis. Below an example *lilkutubi* "books" :

lilkutubi⇔ [kitAb-1 POS: N l+ Al+ +PL +GEN]
li_l_kutub-i⇔
[lemma-ID NOUN PREP+DET+ (plustem) + Genitive]

Although the lexicon is an open linguistic resource, the procedure for updating it is complex. For instance, adding a new verb is an intricate operation. First, the A and C lexica are composed of 561 and 989 entries. Although the two disjoint sets of inflectional and agglutination suffix morphemes are clearly defined in Arabic, the *[prefixes] [stem][suffixes]* representation does not allow two suffix subsets to be defined. Second, the stem lexicon entries corresponding to a lemma are numerous and need to be subcategorized. In other words, a lemma is unfolded into many stems, and one uses a cumbersome subcategorization which mixes up inflectional and agglutinative features of verb stems in order to match with 3 compatibility tables, composed respectively of 2050, 1660, 1200 entries. Such composite data are complex and not transparent for Arabic linguists.
Mesfar (2008) adopts a "lemma-based lexicon" and FSTs for inflection. The project claims 10 000 verb lemmas. The framework is similar to ours since it resorts to classical techniques of lexicon compression and lookup in a full list of inflected -forms. The project does not use root-and-pattern representation. As far as we know, no figures on testing and evaluating the systems are available. The lemma lexicon is wordy such as the extract of the lexicon from Mesfar (2008):

```
ضَرَبَ,V+Tr+FLX=Vdaraba1+DRV=N_daraba1:FlxDRV+DRV=A_daraba1:FlxDRV
# le verbe "ذَكَرَ" et "كَتَبَ" se conjuguent et se dérivent selon les même modèles
ذَكَرَ,V+Tr+FLX=Vdakara2+DRV=N_dakara2:FlxDRV+DRV=A_dakara2:FlxDRV
كَتَبَ,V+Tr+FLX=Vdakara2+DRV=N_dakara2:FlxDRV+DRV=A_dakara2:FlxDRV.
```

FST are difficult to read and maintain (Mesfar, 2006, page 3):

> " أَلَّمَ ", V+Tr+FLX [8] = V_kallama (kallama – *to speak with someone*)
> Among the 122 inflectional transformations which are described in the flexional paradigm "V_kallama", here is one: (<LW> يُ <R4><S> <R><S> /ُ A+P+3+m+s). This NooJ transformation means: position the cursor (|) at the beginning of the form(<LW>) (|kallama), insert "يُ" (yu) into the head of the form (yu|kallama), skip four letters (<R4>) (yukall|ama), erase a letter (<S>) (yukall|ma), insert the vowel "ِ " (i) (yukalli|ma), skip a letter (<R>) (yukallim|a), delete of the following letter (<S>) (yukallim|)and finally insert the final vowel "ُ " (u) (yukallimu|)."

For their morpho-phonological system and in addition to concatenative rules, Carnegie Melon Univ. uses transformational rules to describe alternation of root letters (Cavalli-Sforza, et al., 2005). As far as we know, no figures on lexical coverage or evaluation are available.

The SARF project (Al-Bawab et al., 1994, http://sourceforge.net/projects/sarf/) is based on root-and-pattern representation. Starting from three- and four-consonant roots, it can generate Arabic verbs, derivative nouns, and gerunds, and inflect them. It has over 20 000 verb lemmas. The project uses conventional programming techniques with the Java language and roots encoded in XML files. It uses transformational rules in order to handle alternation of root letters in the Java programs. The patterns are hard-coded in the form of Java code. This work has the advantage of being clearly built on a strong linguistic basis that is the standard morphology in Arabic. However, it neither includes the use of a test collection nor reports a success rate; in addition, updating and correcting a language resource included in source code is complex since it involves two expertises: an Arabic linguist and a programmer; updating data and updating source code obey to different professional practices.

At Université de Lyon 2, the DIINAR project (Dichy & Ferghali, 2004) was developed for terminological and translation purposes. DIINAR.1 includes a total number of 119,693 lemmas, fully vowelled, among which 19,457 verb lemmas. A conventional programming framework and databases are used for generation and analysis with a lemma-based lexicon encoded according to this framework. As far as we know, no figures on testing and evaluating the system for morphological annotation are available.

For a complete survey of morphological parsers, readers should consider Al-Sughaiyer & Al-Kharashi (2004) and Habash (2010).

3. Method of description

3.1 A taxonomy for verb inflection

Our method is based on a precompiled diacriticized full-form dictionary with all possible inflected forms and their orthographic variations due to morphophonemic alternations. We exclude from this inflectional representation agglutinated prefixes and suffixes such as conjunctions and pronouns. We associate morphosyntactic feature values to each entry in the generated list of 2.43 million surface forms. In order to obtain this list, we provide a list of lemmas manually associated to codes defined by a taxonomy, each code representing a transducer. The full-form list is produced after inflecting each lemma by applying the encoded transducer (Silberztein, 1998).

Arabic and other Semitic languages have long been described in terms of a *root* interwoven with a *pattern*. The root is a sequence of consonants. Each Arabic verb contains 3 or 4 consonants that remain generally unchanged in all conjugated forms and make up the consonantal root; all the remaining information on a conjugated form is called 'pattern'. For example, *yakotubuwna = [ktb & ya1o2u3uwna]* is obtained through the interdigitation of the root *ktb* with the pattern of active-Perfect-3person-masculine-plural-indicative *ya1o2u3uwna*. Below some precisions:

- Some root consonants change. They are the glottal stop, noted *h* in the taxonomy, and glides, noted *w*, *y*; those that never change are written in patterns in the form of their position 1, 2, 3 or 4.
- At the surface level, the orthographic representation of glottal stop and glides can change. The glottal stop is represented by six allographs depending on the context. At phonological level, the glides become short vowels */ i, u/* or long vowels */a:*, *i:*, *u:/* or are omitted and transcripted as *zero-vowel*, *o*[3] (see also footnote 4).
- A pattern indicates the position of its letters relative to the root consonants. Generally, these letters are vowels and/or affixes related to derived verb form such as *IisotakotabuwA = [ktb & Iisota1o2a3uwA]*. The surface form may also be subdivided in *[prefix] [stem] [sufix]*. The *stem pattern* formalizes all infixation operations such as *kotub = [ktb & 1o2u3]*. Inflectional prefixes and suffixes can be concatenated subsequently to the stem form y*akotubuwna = [ya] [ktb & 1o2u3] [uwna]*.
- The third root consonant can be identical to the second one. In the root, it is represented by a gemination mark *G*, and in the pattern, by *2*, such as *madadota = [mdG & 1a2a2ota]*.
- By convention, the perfect-3rd person-masculine -singular is the form used as lemma. The corresponding pattern is called the canonical pattern. All patterns are defined in function of the canonical pattern.

Verbal pattern classes are clearly defined in Arabic grammar but root-classes are intricate and involve a complex terminology. Root-classes are defined according to the nature of some of the root consonants: regular, weak, geminated, with glottal stop, and to their position 1, 2, 3 or 4. In this terminology, *qaAla/yaquwlu* قال "say" is a *hollow verb of w kind*, with a weak consonant *w* at the second position; whereas *baAEa/yabiyEu* باع "sell" is a *hollow verb of y kind*. Moreover, two or three special values of the root consonants can appear at the same time. A verb like *OataY/yaOotiy* أتى "arrive" has a glottal stop at the first position and a weak consonant *y* at the third position. A classification with nature/position criteria and each with 4 sub-criteria yields to an intricate terminology and is not

3 The zero-vowel marks the absence of vowel between two consonants.

consensual in Arabic grammar.
Our classification is bi-dimensional like the traditional one and based on the traditional pattern-classes which are reused and root-classes which are redefined more simply. Traditional grammar defines an inflectional verbal class by a pattern-class and a root-class. Triliteral verbs are compatible with 16 possible canonical patterns and quadrilateral verbs with 4 canonical patterns. Our classification defines 31 root-classes. The root classes are defined according to the nature of the root consonants. The special values for the consonants are *w, y* and the glottal stop (h). An irregular root is a root with at least one special value in its consonants. The inflected forms of a verb are easily predictable on the basis of the features of the root. We revisited and simplified, with no loss of information, the root-based traditional classification by using three consonantic slots, noted *123*, except for special values: glottal stop (*h*), *w*, *y*, for each slot; and when the 3rd root consonant is identical to the 2nd, the slots are noted *122*. Thereby, the lemma *ktb* will be encoded *$V3au-123* where:

$ is the Semitic mode for FST which means the root consonants interdigitate into the pattern: *[ktb & ya1o2u3u] = yakotubu;*
V is the verbal POS;
3au is the class of triliteral verbs used with the patterns *1a2a3/ya1o2u3* for perfect/ imperfect;
123 is the class of roots in which no slot is occupied by a special value.

Each root/canonical-pattern pair corresponds to a lemma. This representation seems well-founded and also well-established in Arabic morphology. Above all, it is ubiquitous in the Arabic-speaking world. Below, some examples from the lexicon:

```
/Lemma,encoding/ canonical-patt. Special values
-----------------------------------------------
/ simple forms
نقش,$V3au-123 / 1a2a3a/ya1o2u3u no special values
جرّ,$V3au-122 /    third root identical to second
عاد,$V3au-1w3 /      with waw as a second root
غفا,$V3au-12w /      with waw as a third root
فتح,$V3aa-123/ 1a2a3a/ya1o2a3u
لمز,$V3ai-123 / 1a2a3a/ya1o2i1u
حاك,$V3ai-1y3 /   with yeh as a second root
سرى,$V3ai-12y /   with yeh as a third root
أوى,$V3ai-hwy /   with hamza, waw and yeh
علم,$V3ia-123 / 1a2i3a/ya1o2a3u
وطئ,$V3ia-w2h /   waw and hamza as 1rst and 3rd
كرُم,$V3uu-123 / 1a2u3a/ya1o2u3u
حسب,$V3ii-123 / 1a2i3a/ya1o2i3u
/ Derived forms
أقبل,$V61-123  / Aa1o2a3a
دشّن,$V62-123  / 1a2Ga3a
داهم,$V63-123  / 1aA2a3a
إنشغل,$V64-123 / Iino1a2a3a
إنطلى,$V64-12y /   with yeh as a third root
إختنق,$V65-123 / Ii1ota2a3a
إزهرّ,$V66-123  / Ii1o2a3Ga
تهاجن,$V67-123 / ta1aA2a3a
تآكل,$V67-h23  /   with  hamza as a first root
تحدّد,$V68-122  / ta1a2Ga2a with identical 3rd root
تلكّأ,$V68-12h  /   with hamza as a third root
إستبسل,$V69-123 / Iisota12a3a
اعشوشب,$V70-123 / Ii1o2aw2a3a
/ Quadriliteral roots
بعثر,$V40-1234  / 1a2o3a4a a quadriliteral root
طمأن,$V40-12h4  /   with hamza as a third root
دمدم,$V40-1212  /   a geminated quadriliteral root
تبعثر,$V41-1234 / ta1a2o3a4a
تلألأ,$V41-1h1h  /    a geminated root with 2 hamzas
```

Below, some of the 31 possible combinations of root-classes related to class-pattern V3ia. Some root-classes are empty which means that there is no verb with such root-classes for class-pattern V3ia:

```
/Lemma,encoding/  /lemma-transliteration
علم ,$V3ia-123   /Elm
ظل  ,$V3ia-122   /ZlG
أمَ  ,$V3ia-h22   /OmG
ألف ,$V3ia-h23   /Olf
رئف ,$V3ia-1h3   /ref
ظمئ ,$V3ia-12h   /Zme
//First weak root consonant
ودَ  ,$V3ia-w22   /wdG
    ,$V3ia-wh3
وطىء,$V3ia-w2h   /wTe
وجع ,$V3ia-w23   /wjE
    ,$V3ia-y22
يئس ,$V3ia-yh3   /yes
يقظ ,$V3ia-y23   /yqZ
```

The format of the lexicon is a list of lemma entries. In our format, the string before comma transcribes plain letters and the gemination mark but no short vowel diacritics. The pattern includes the encoding of short vowels (*a, i, u).* This transcript choice is consistent with usual practice in traditional paper dictionaries.
Our full-form lexicon is produced by FSTs. The FST output format is *surface-form,lemma.V:feature-values* such as *:*
كتب,تَكْتُبُ.V:aI3fsN
/active-Imperfect-3rdpers-fem-sing-iNdicative

The *feature values* are :
- Voice: active (a), passive (b);
- Tense: Perfect, Imperfect, Imperative (Y);
- Person: 1, 2, 3;
- Gender: masculine, feminine;
- Number: singular, dual, plural;
- Mode: indicative (N), Subjunctive, Jussive, Energetic.

In the following two sub-sections, we present first inflectional transducers and then inflection-related orthographic adjustments.

3.2 The inflection transducers

An inflection transducer specifies the inflectional variations of a word. It is shared by the class of words that inflect in the same way. The input parts of the transducer encode the modifications that have to be applied to the canonical forms. The corresponding output parts contain the codes for the inflectional features. A transducer is represented by a graph and can include subgraphs. The transducers are displayed in Unitex style, i.e. input parts are displayed in the nodes, and output parts below the nodes.

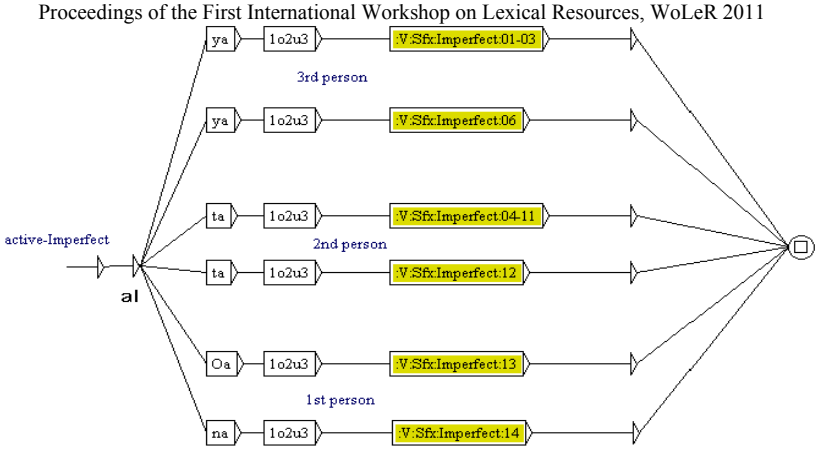


Fig 1. The active imperfect (aI) subgraph. Each path contains a prefix, a stem-pattern and a subgraph of suffixes. The Person-Gender-Number variations are numbered from 01 to 14.

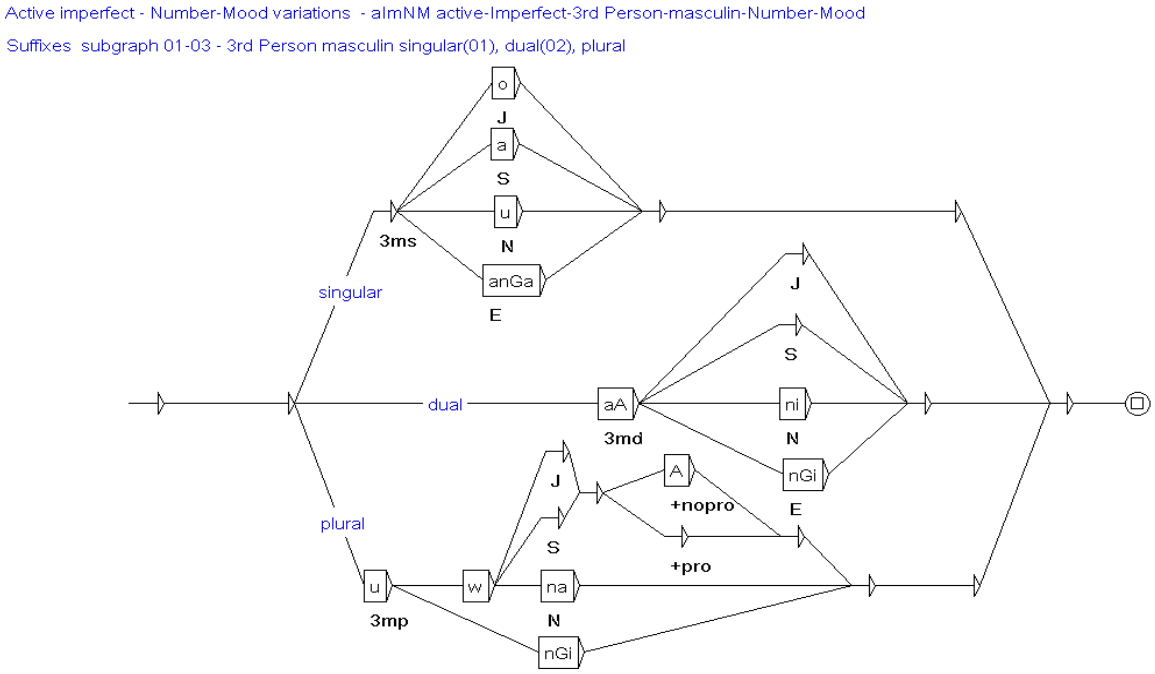


Fig 2. The 01-03 subgraph represents Number-Mode suffix variations for active Imperfect 3rd Person masculine, related to Person-Gender-Number-Mode variations.

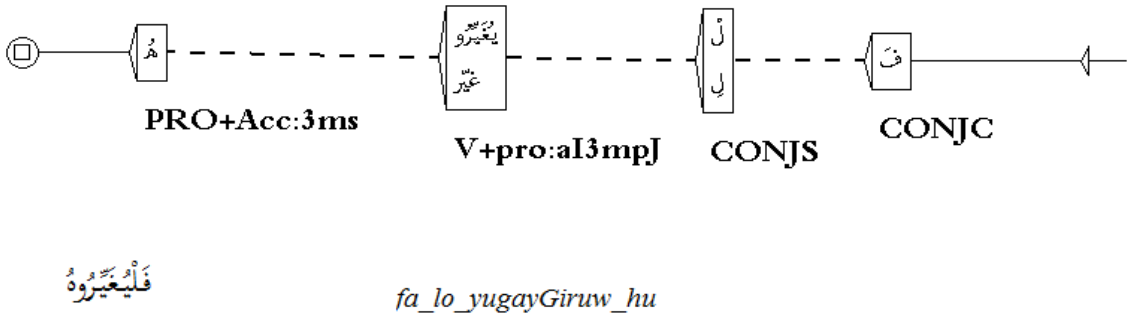


Fig 3. Text automaton as output of the application of a graph dictionary. Here a morphological analysis of *faloyugayGirohu (and_to_change-they_it)*. The morphological dictionary graph restricts the selection to V+pro agglutinated variant only. Dashed lines connect segments in the same token.

A Buckwalter transliteration is used as a standard to map Arabic characters into Latin ones. An XML version of this transliteration was created in order to handle this format. We create a modified version of the XML version where all special characters such as (', |, * , $, ~) are respectively replaced by (c, C, J, M, G) [4]. Many systems use special characters in a special way.
In order to generate the full-form dictionary, the following steps are accomplished.
- The lemma lexicon is transliterated.
- The FSTs are applied to the list and produces a transliterated full-form dictionary output.
- The output is transliterated into Arabic script.
So, both the lemma lexicon and the full-form dictionary are in Arabic script which is handier to read for Arabic linguists.

For example, the lexical entry `ktb,$V3au-123` is processed by the transducer named *V3au-123* in order to get all inflected forms. The main graph contains five subgraphs referring to the five voice-tense variations. In turn, each subgraph (Fig. 1) contains suffixes of Person, Gender, Number for the perfect and Person, Gender, Number, Mode for the Imperfect (Fig. 2).

3.3 Inflection-related adjustments

The inflectional taxonomy takes into account variations due to orthographic adjustment and morphophonemic assimilations. The phoneme involved in the variation is replaced by a gemination mark or by another phoneme. At morpheme boundaries between a stem and a suffix, the first letter *n* and *t* of the perfect suffix is changed to gemination mark like in *daxGan+naA => daxGanGaA* , "smoked-we"; *Oavobat + tu => OavobatGu* "demonstrated-I". Our taxonomy includes the inflectional classes Vpp-12n, Vpp-12t in order to take into account such phenomena. In our resource, we have counted 614 entries in Vpp-12n and 154 in Vpp-12t root-classes.

Due to morphophonemic variations, the *t* in the canonical pattern V65 or *Ii1ota2a3a* (اِفْتَعَلَ) has an orthographic variation depending on the value of the first root consonant. It is replaced by emphatic *T*, or *d*, or by gemination mark *G*. The subclasses V65T, V65d, V65G encode the *t* variation, we have counted: 46 entries with V65T-rrr such as `ISTfY,$V65T-12y` إصطفى; 31 entries with V65d-rrr such as `Izdwj,$V65d-1w3` إزدوج; and 114 entries with V65G-rrr such as `ItGbE,$V65G-123` إتّبع or `ItGSl,$V65G-w23` إتّصل.

4. Agglutination and omission of diacritic

4.1 Orthographic adjustments and agglutination

In Arabic, a token delimited by spaces or punctuation symbols is composed of a sequence of segments. Each segment in a token is a morpheme. In Unitex, this segmentation is formalized via a morphological dictionary graph. Such graphs introduce morphological analyses in the text automaton (Fig 3) where dashed lines connect segments.

The combination of a sequence of morphemes obeys a number of constraints. Checking these constraints is necessary to discard wrong segmentations. In Arabic, a verbal token is composed by one morpheme <*V*> or the concatenation of up to 4 morphemes such as:
<*CONJC*> <*CONJS*> <*V*> <*PRO+accusative*>
where <CONJC> is a coordinating conjunction, <CONJS> is a subordinating conjunction and <*PRO+accusative*> an agglutinated object pronoun.

<*CONJC*> combines freely with any inflected verb. The <*CONJS*> constraints the verb to the Imperfect Subjunctive or Jussive. Finally, an inflected verb form is often insensitive to the agglutinated pronoun but some forms are sensitive like forms with a glottal stop as the third root consonant.

The subgraph selects only V+pro variants from the full-form dictionary (cf. Fig 3). When followed by a pronoun, a verbal segment may have an orthographic adjustment. This is often the case when the verbal segment ends with a long */a:/ A*, its allograph *Y*, or a glottal stop which has 6 allographs depending on its position and the surrounding vowels. For verbs, the roots with a glottal stop as the third consonant change their graphemic representation. A suffix subgraph related to classes Vpp-rrh represents the orthographic variations of an ending glottal stop due to pronoun agglutination.

The generation of the agglutinable variants of an inflected verb is performed directly with a lexicon of words, which is another way to implement a rule. In fact, the dictionary graph links each morphological variant to the correct context, which also expresses a rule. The variants are generated during the compilation of the resources, not at analysis time as in rule-based systems in which a rule should compute each morphological variant at run time, then link each variant to the correct context. The advantage of our method is that it simplifies and speeds up the process of annotation.

4.2 Diacritics

Diacritics are often omitted in Arabic written text. According to our corpus study of 6930 tokens from Annahar newspaper, 209 tokens (3%) include at least a diacritic. 140 tokens (2 %) are with the *F* diacritic *(–an)* and 57 (1 %) are with gemination mark *G*, in which nearly 0.8 % is related to a verbal form. 9 are with the short vowel *u*. For the *u* diacritic, 7/9 involve a passive verbal form. For the gemination diacritic, 49/57 involve a verbal form and are the following.
- 41 to V62 refer to *1a2**G**a3a* derived form (فَعَّلَ).
- 5 to V68 refer to *ta1a2**G**a3a* derived form (تَفَعَّلَ).
- 2 to V65G refer to *Ii1**G**a2a3a* derived form (اِفْتَعَلَ).
- 1 to V3au refers to *ya1o2**u**lu* a triliteral simple form (فَعَل يَفعُل).

4 The Transliteration in Unitex Arabic <=> Latin: ء, c; أ, C; أ, O; ؤ, W; إ, I; ئ, e; ا, A; ب, B; ة, P; ت, T; ث, V; ج, J; ح, H; خ, X; د, d; ذ, J; ر, r; ز, z; س, s; ش, M; ص, S; ض, D; ط, T; ظ, Z; ع, E; غ, g; ف, f; ق, q; ك, k; ل, l; م, m; ن, n; ه, h; و, w; ى, Y; ي, y; ً, F; ّ , N; ِ , K; َ, a; ُ, u; ٍ, i; ٌ, G; ْ, o;

Editors generally display diacritics for unusual forms such as passive verb forms. When some are displayed, they can avoid misinterpretations to the reader. For verbs, diacritics are the short vowels *(a, i, u)* or the gemination mark followed by a short vowel. Arabic verbs can include a sequence of two diacritics: the gemination mark followed by a short vowel. In the case of two diacritics, diacritics omission is not totally free. One can omit the two diacritics or the last diacritic but never the gemination mark alone.

Consequently, processing written Arabic text should take into account undiacriticized and partially diacriticized text. A lookup procedure in Unitex[5] has been adjusted to deal with omission of diacritics in Arabic. This procedure finds in the diacriticized full-form dictionary all possible diacriticized candidate forms compatible with a given undiacriticized or partially diacriticized form. When a diacritic is present in a surface form, the lookup procedure excludes the candidates in the lexicon which do not have that diacritic at the same position.

5. Some figures

Our lexicon is composed of 15 400 entries. Each entry is inflected into 144 surface forms and in average 158 forms if we include orthographic variations due to agglutination. The size of the full-form dictionary is 2.43 million surface forms. The size of the full-form dictionary in plain text is 132 Megabytes in Unicode little Endian and is compressed and minimized into 4 Megabytes which is loaded to memory for fast retrieval. The generation, compression and minimization of the full-form lexicon lasts two minutes[6] on a Windows laptop.
The number of main inflectional graphs is 460. Each main graph is composed of 5 subgraphs for voice-tense features variations, that is 2300 subgraphs. These subgraphs use also 540 suffix subgraphs related to person-gender-number-mode features. In all, the number of graphs and subgraphs is 3300 (460+2300+540), to be compared with nearly 100 graphs and subgraphs dedicated to the verbal inflection system for Brazilian Portuguese constructed also for Unitex (Muniz et al. 2005). A sample will be freely available from the time of the workshop.

We have noticed that many simple triliteral verbs may have orthographical variants related to the variation of the vowel after the second root consonant. However, these variations may correspond to meaning differences; therefore we should have different entries. In order to facilitate the encoding scheme, all orthographic variants of verbs are encoded in separate entries. In our lexicon, a verb may have several inflectional codes. These codes can correspond to different lexical items or to orthographic variants of the same item. In the future, we plan to encode different lemmas if the different inflectional behaviours are correlated to differences at other levels, e.g. semantic, which is the case of `Hsb,$V3au-123` "count", and `Hsb, $V3ii-123` "think". One should also encode a single lemma if the inflectional behaviours are a free variation, such as for `kfl,$V3au-123` and `kfl,$V3ai-123` "grant". Out of a total 4135 simple triliteral root in the lexicon, 1278 triliteral root have several inflection al codes.

Some inflectional classes are redundant such as V62-122, which is identical to V62-123, whereas V65-122 is different from V65-123. In order to make the encoding scheme easier to handle for Arabic linguists, we have duplicated the inflectional graph V62-122. The 122 root-class delimits two classes in nearly all other cases. We estimate such redundancy at 15%. We offer a simple encoding scheme with duplicated inflectional classes in order to make it unnecessary for Arabic linguists to memorize in which cases some features have to be marked.

6. Evaluation

We have chosen the NEMLAR Arabic Written Corpus (Attia et al., 2005), first to improve our lexicon of verbs, and then to constitute our test collection. The Nemlar data consists of about 500 thousand words of Arabic text from 13 different genres. The text is provided in 4 different versions: raw text, fully diacriticized text, text with Arabic lexical analysis, and text with Arabic POS-tags. The database was produced and annotated by RDI, Egypt, for the Nemlar Consortium.
The extraction of occurrences of verbs from "text with Arabic POS-tags" provided 50 000 occurrences of verbs. These occurrences were split in two disjoint parts: nearly 40 000 token occurrences (11050 token types) for correcting the resource and a test collection of 10 000 token occurrences (5222 token types) for testing it after the correction stage.
The test collection shows that 10 verbs lemmas were missing in our lexicon[7]. Hence, the fault rate of the resource is 0.1% in this corpus. Let us assume that a page is composed of 50 lines/page, 10 tokens/line, 1 verb/10 tokens. In other words, in 20 pages of real corpus, our resource fails to recognize 1 verb.
In order to compare our lexicon with the Buckwalter resource, we ran BAMA on the first 550 occurrences of verbs of the same test collection. 14 occurrences of verbs were unrecognized, which represents a 2.5 % error rate, i.e. 25 times the error rate of our resource. The unrecognized tokens involve: 10 missing passive stems, 2 imperative stems and 2 missing verb lemmas.
Morphosyntactic tagging is generally part of a pipeline of written text processing. In a common undiacriticized Arabic corpus, most verbs have two possible analyses, one as active and one as passive. The lack of passive stems in the Buckwalter resource leads to assign only the active tag to verbs, which can jeopardize a subsequent deep syntactic parsing of a sentence.
A fallback procedure in order to assign morphosyntactic

5 The lookup procedure was adjusted by Sébastien Paumier.

6 At Columbia University, MAGEAD Project constructs an Arabic resource according to Buckwalter's Prefixes-Stem-Suffixes representation. They describe an Arabic lexicon based on root-and-pattern representation and rules dedicated to orthographic variations due morphophonemic alternations; and other rules dedicated to orthographic adjustment due to agglutinations (Habash & Rambow, 2006). The program needs more than 15 hours to generate such resource (Owen Rambow, personal communication).

7 jzm,$V32-123; qrGZ,$V62-123; thrGb,$V68-123; rDb,$V33-123; kfl,$V34-123; tnAqM,$V67-123; sAb,$V32-1y3; zEq,$V33-123; DnG,$V32-1nn; tAh,$V32-1y3

features to unrecognized tokens is often included in a language processing pipeline. Since our fault rate is 0.1 %, it might be useless to construct a fallback procedure for unrecognized verbs when this resource is used.

7. A conclusion and perspectives

We elaborated a model for Arabic verbs with the following features. A detailed and simple taxonomy is based on Semitic morphology. Lemma-based verbs are used as entries in the lexicon. FSTs are used to produce inflected forms. Agglutination is described independently from inflection.Our experimentation shows that the method outperforms state-of-the-art systems of Arabic morphological annotation.

We made language resources the central point of the problem. All complex operations were integrated among resource management operations. The output of our system is accurate and informative; the language resources used by the system can be easily updated by an expert of Arabic independently from computational linguistics experts, which allows users to control the evolution of the accuracy of the system. Morphological annotation of Arabic text is performed directly with a lexicon of words and without morphological rules, which simplifies and speeds up the process. The undiacriticized, partially and fully diacriticized Arabic text can be annotated excluding incompatible analyses.

We reuse traditional Semitic patterns and we provide a clear scheme for root-class encoding by avoiding intricate terms. Root-and-pattern representation facilitates our task in encoding the lexicon since it is a standard but also it helps to debug our transducers quickly which is not the case of a rule-based system.

This work opens several perspectives. The resources can be extended by running the annotator and analysing output. Another perspective is to extend this methodology to inflection of noun and adjective, mainly to encode singular and the plural under the same lemma entry using Semitic patterns فَـعِـيل فُـعَـلاء. For example, the pair *raeiys*, *ruWasaAc* (رَئِيس رُؤَسَاء) "president" will be represented by one entry:

```
raeiys,$N3_1a2iy3-1u2a3Ac-1h3
nabiyl,$N3_1a2iy3-1u2a3Ac-123
```

where number *3* denotes a triliteral root; *1a2iy3-1u2a3aAc* is a pattern pair that represents singular-plural variations; and *1h3* (vs *123*) encode the glottal stop variations of the 2nd consonant root *(e =>W)*.

8. References

Al-Bawab, M., Mrayati, M., Alam, Y.M., Al-Tayyan, M.H. (1994). A computerized morpho-syntactic system of Arabic. In *The Arabian Journal of Science and Engineering*, 19, 461-480. Published by KFUPM, Saudi Arabia.

Attia., M., Yaseen., M., Choukri., K. (2005). Specifications of the Arabic Written Corpus produced within the NEMLAR project, www.NEMLAR.org.

Beesley, Kenneth R. (1996). Arabic finite state morphological analysis and generation. *In COLING'96, volume 1*, pages 89– 94, Copenhagen, August 5-9. Center for Sprogteknologi. The 16th International Conference on Computational Linguistics, 1996.

Buckwalter, T. (2004). Issues in Arabic Orthography and Morphology Analysis. *In Proceedings of the COLING 2004*. Workshop on Computational Approaches to Arabic Script-based Languages, pages 31–34.

Buckwalter Arabic Morphological Analyzer Version 1.0. (2002). LDC Catalog No.: LDC2002349.

Cavalli-Sforza, Soudi, Mitamura (2000). Arabic Morphology Generation Using a Concatenative Strategy. In *Proceedings of the 6th Applied Natural Language ProcessingConference (ANLP 2000)*, pages 86–93, Seattle, Washington, USA.

Dichy, J., Farghaly, A. (2003). Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: there what basis should is multilingual lexical database centred on Arabic be built? *In Workshop on Machine Translation for Semitic Languages*, New Orleans, USA.

Habash, N., Rambow, O. (2005). Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. *In Proceedings of the Conference of American Association for Computational Linguistics* (ACL05).

Habash, N., Rambow, O. (2006). MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 681–688, Sydney, Australia, July.*

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypoll Publishers.

Huh, H.-G. Laporte É. (2005). A resource-based Korean morphological annotation system. *In Proc. Int. Joint Conf. on Natural Language Processing*, Jeju, Korea, 2005.

Kiraz, A. (2004): http://www.scribd.com/doc/46443095/Computational-Nonlinear-Morphology-With-Emphasis-on-Semitic-Languages-Studies-in-Natural-Language-Processing-9780521631969-41686

Mesfar, S. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. Thèse, novembre 2008, Université de Franche-Comté.

Mesfar, Slim. (2006). Standard Arabic formalization and linguistic platform for its analysis in *Proceedings of Arabic NLP/MT conference*, London, 2006

Marcelo C.M. Muniz, Maria das Graças V. Nunes, and Éric Laporte (2005). UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. *Workshop TIL'05*. pp. 2059–2068.

Paumier, Sébastien. (2011). *Unitex - manuel d'utilisation 2.1* , University of Marne-la-Vallée.

Silberztein, Max. (1998). INTEX: An integrated FST toolbox, in Derick WOOD, Sheng YU (éd.), *Automata Implementation*, p. 185-197, Lecture Notes in Computer Science, vol. 1436. Second International Workshop on Implementing Automata, Berlin/Heidelberg: Springer.

Smrz, Otakar. (2007). ElixirFM — Implementation of Functional Arabic Morphology. *In Computational Approaches to Semitic Languages*, ACL 2007, Prague.

Al-Sughaiyer, Imad A., Al-Kharashi, Ibrahim A. (2004). Arabic Morphological Analysis Techniques: A Comprehensive Survey. *In Journal of the American Society for Information Science and Technology*, 55(3):189–213.

ARTES: an online lexical database for research and teaching in specialized translation and communication

Mojca Pecman, Natalie Kübler
CLILLAC-ARP EA 3967
Université Paris Diderot – Paris 7
5 rue Thomas Mann
75205 Paris Cedex 13
PRES Paris Sorbonne Cité
E-mail: mpecman@eila.univ-paris-diderot.fr, nkubler@ eila.univ-paris-diderot.fr

Abstract

This paper presents the ARTES database (Aide à la Redaction de TExtes Scientifiques / Dictionary-assisted writing tool for scientific communication) and the underlying approaches to lexicography, terminology, languages for special purposes (LSPs), and lexical resources creation, behind the design of the database. This new type of lexical resource has been developed within the ARTES project, whose main objective is to explore the interaction between research and teaching in the areas of applied linguistics such as specialised translation and LSP communication. As a multilingual multidomain language resource targeting various LSP users – students, translators, experts, subject specialists, teachers, linguists –, the ARTES database offers a comprehensive approach to lexical resources: terminological, phraseological, domain-specific, domain-free, semasiological and ultimately onomasiological. The underlying research orientations are thus broad and allow to investigate various language mechanisms which operate on lexico-discursive level, and consequently to fine tune the database in order to take into account these various linguistic phenomena.

1. Introduction

The present paper is an overall study of the interactions between research and teaching in the domain of lexical resource creation that have been taking place within the ARTES project[1]. Launched in 2007 at Paris Diderot University - in the frame of the ESIDIS-ARTES scheme - the ARTES project was designed to bridge the gap between research and teaching in a number of related areas: terminology, phraseology, translation, LSPs, lexical resources, corpus linguistics, genre and discourse analysis. It enables us to tackle some core research problems related to the conceptual design of lexical resources which seek to integrate language phenomena currently observed through corpus analysis and on the linguistic levels of terminology, phraseology and discourse.

In 2010, a major tool was added to the project, an online database, opening up new avenues of research and facilitating the creation of lexical resources and development of dictionaries to meet the specific needs of speakers using LSPs: researchers, experts, translators, students, teachers. In addition to terminology, the database highlights phraseology, whether domain-specific or domain-free.

Although comparable to some extent to terminological databanks such as Termium[2], Grand Dictionnaire Terminologique[3] or Eurodicautom, now known as IATE[4], the ARTES database is closer to initiatives on lexical resources creation where teaching, research and database development are very closely related, such as DiCoInfo[5] database, and the latest DiCoEnviro[6] (L'Homme 2007) or WebTerm[7], a project of the Institute for Information Management in Cologne.

After these preliminary remarks, we shall first discuss the general approaches to specialised lexical resources adopted in ARTES database. The turning of the database into an online electronic dictionary by providing an interface for data access will be exemplified in the second part of the paper. In the last part, we evaluate the relevance of research conducted on terminology, phraseology and specialised discourse for a fine tuning of database architecture.

2. Creating lexical resources with ARTES database

The present section is an overview of the general scheme of the ARTES project and its goals, and of the ARTES database architecture designed to host LSP resources.

2.1 Presentation of ARTES project

ARTES is an ambitious and innovative project developed with the aim of bridging the gap between research and teaching in LSP translation and communication. It is carried out at the Paris Diderot University by a group of researchers working on LSP, Corpus Linguistics and Translation Studies: Kübler,

[1] ARTES project homepage: http://www.artes.univ-paris-diderot.fr

[2] The Government of Canada's terminology and linguistic databank: http://www.termiumplus.gc.ca

[3] Dictionary of the Office québécois de la langue française : http://www.granddictionnaire.com

[4] InterActive Terminology for Europe database: http://iate.europa.eu

[5] Dictionnaire fondamental de l'informatique et de l'internet : http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi

[6] Dictionnaire fondamental de l'environnement : http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi

[7] http://www.iim.fh-koeln.de/webterm/webtermsamm_e.htm

Pecman & Bordet 2011; Froeliger 2008; Humbley 2008; Kübler 2011; Kübler & Pecman forthcoming; Pecman 2005, 2008; Pecman *et al.* 2010; Volanschi et Kübler 2011. The target was to construct a model for collecting lexical resources in LSPs which would be flexible enough to allow developments in line with advances in research and changing learning needs. After testing several experimental models, we decided on SQL database technology with online applications for editing and retrieving data.

Several earlier developments paved the way for acquiring knowledge necessary for designing this online database, namely BasTet[8] designed in 2006 by Claudie Juilliard, Terminom1[9] developed first in 2004 by Kübler and Juilliard and redeveloped in 2007 by Kübler and Pecman, and the LangYeast[10] combinatory dictionary developed by Volanschi (2008).

The architecture of the ARTES database is inspired by BasTet which was first developed using Microsoft Access database management system. In ARTES, new functions were added based on a better understanding of General Scientific Language (GSL) and the processing of domain-free phraseology (Pecman 2004, 2007, 2008). The transfer of technology from Access DBMS to an online SQL database was entrusted to (e)Kudji company. The new online database is currently under development but the main stages of transfer were achieved by November 2010.

Although not an XML database, The ARTES database was developed in agreement with TBX and TMF standards for terminological databases. The structure of a terminological entry, the specific data categories and the relational disposition of data adopted in ARTES scheme are very close to meta model of ISO 16642 standards and guidelines for creating terminological data collections. Nevertheless, in the present state of the tool, the terminological data collection (TDC) scheme recommended by ISO 16642 for providing information on concepts of specific subject fields is not yet integrated. In the future developments of ARTES DB, we intend though to provide a preliminary conceptual level analysis where concepts would act as pivots ensuring linguistic transfer between different languages, and consequently an access to data through ontologies.

2.2 General architecture of ARTES database

The ARTES database is a relational database designed to contain lexical information spread over a number of tables. Some aspects of the ARTES database architecture were already discussed in Pecman *et al.* (2010), and Kübler & Pecman (forthcoming). There are some forty tables in the ARTES database, which can be divided into three subgroups according to the type of information stored in the tables: data tables, labelling tables, and relational tables.

Data tables are central tables in which all key information is stored. There are six data tables in ARTES containing respectively sources, terms, contexts, definitions, specific collocations, generic collocations, and notes. The table for sources serves to record bibliographical references of textual or oral sources used for referencing definitions, contexts or notes. In the table for terms, all terms are recorded no matter what language or domain. The specifications on the language and domain are given through descriptors provided by labelling tables. Context tables serve to record contexts taken from various sources and which serve as examples for illustrating the usage of terms or collocational phenomena. The tables for definitions contain all the definitions of the terms recorded in the database. For each term, all relevant existing definitions are provided and, if necessary, a new definition is drafted. The table of specific collocations is designed to record the most frequent collocations associated with terminology. In this way, two tables were designed to separate specific collocations related to terms from generic collocations related to discourse functions. The table of generic collocations is designed to list frequent word combinations used in a variety of LSP domains and which are related to discourse functions, the latter being recorded in a label-type table (see hereafter). Finally, the table for notes contains various observations on resources stored in data tables (e.g. an additional commentary on the meaning of the term as a complement to its definition, or an observation on the choice of terms indicated as synonyms).

Labelling tables are the tables with pre-defined values which describe and classify the resources stored in the data tables. The predefined values act as labels or descriptors. They allow us to adopt a descriptive approach to language data. Some labelling tables contain closed-class type values, such as the table of grammatical categories linked to the table of terms. Other labelling tables are open-class tables and can be modified or completed according to the results of research conducted in relation with language resource creation. For instance, the table of discourse functions which offers some eighty classes for categorizing generic collocations according to their general meaning or function in LSP discourses, is an open-class table.

The relational tables are necessary for establishing various links between data (between equivalent terms across languages, between equivalent pairs of collocations, between synonyms within a language, between a hyperonym and its hyponyms, and so on).

It should be mentioned that in the relational database, all tables are eventually linked together in one scheme which forms the architecture of the database: terms are linked to definitions, contexts and specific collocations, which are in turn linked to sources and notes. Furthermore, terms can be linked to other terms to indicate language equivalences or synonym pairs or sets,

[8] http://wall.eila.univ-paris-diderot.fr/bastet (restricted access)

[9] http://terminom1.eila.univ-paris-diderot.fr (restricted access)

[10] http://ytat2.ijm.univ-paris-diderot.fr/LangYeast

and similarly collocations can be linked to other collocations to indicate equivalent pairs or to form semantic synonymous sets, and so on.

3. Designing language resources for LSP communication and translation

In the ARTES project, the LSP communication and translation are tackled from the learner and professional perspective.

3.1 Taking into account teaching needs in LSP communication and translation

The ARTES database was designed to cater for teaching and learning needs in specialised translation at the department of Applied Languages of Paris Diderot University. The students in Master Studies in Specialised Translation and Language Engineering[11] are introduced in the theories, methods and applications of terminology, lexical resource creation, and corpus linguistics, with emphasis on corpus linguistic tools and information retrieval. A combination of these courses allows the students to develop skills and acquire knowledge crucial for achieving a high quality translation of LSP texts. The final result of the interaction of these various disciplines is presented in a form of Master's dissertation. The ARTES database is designed to allow students to participate in the project by creating LSP resources in relation with the text they translate. In turn, the database offers useful functions for teachers to help them follow students' work in progress and evaluate the resources compiled by students. Special effort was made to design the editing and management interface, commonly called the back-office, to anticipate these user situations.

Consequently, a very important feature of the ARTES DB project is that data is compiled mainly, but not exclusively, by LSP and translation learners. The overall methodology used to ensure the quality of data collected consists in three key procedures. The first one is the method itself followed by learners which is based on thorough comparable corpus analysis and an exchange with domain experts, which leads to a design of a domain ontology. The acquired knowledge on the domain, combined with the knowledge on terminology processing, allows the learners to select and process terms and relevant linguistic information adequately. The second procedure consists in the reviewing and validating resources by domain experts. An onging collaboration with experts in Earth and Planetary Sciences form a STEP department[12] and Institut de Physique du Globe de Paris (IPGP)[13] of Paris Diderot University, allows us to apply this procedure efficiently to a number of disciplines. The third and crucial stage in building language resources is the overall normalisation, correction and validation of resources by terminologists, which should be devised in the near future. We hope thus that the overall methodology will yield satisfactory results.

Students in Master's Studies are also invited to question the theoretical and methodological premises on which the description of language data in ARTES is based by testing them against “real life” translation problems.

The data recorded in the database can be retrieved via an online application specifically designed to take into account various LSP communication contexts.

3.2 Designing an online electronic dictionary for LSP users

The ARTES project had led to the design of an online terminological and phraseological database for storing and managing structure-rich information with the possibility for multiple criteria and multiple-level query. The database is searchable through a database application accessible online[14]. The access to data was devised with special care to targeted users: translators, teachers, students, domain experts and linguists. The intention behind the design of the interface for data access was to explore the possibilities for providing a dictionary-assisted writing tool for scientific communication. In the choice of the name for this application, the priority was given to target users which are largest in number: science students and experts who need to write articles or other text types in their second language within the scope of their discipline. It turns out that these are the most numerous users among students and researchers of Paris Diderot university which is a large multidisciplinary university hosting departments and research centres in Humanities, Sciences and Medicine.

The ARTES database has two interfaces: one for editing and management purposes and one for retrieving information and displaying it in the form of an LSP dictionary. The latter interface has been designed to display data recorded in the database functionally, following Leroyer's approach to functional lexicography according to which *development of a dictionary is determined by users needs and made to serve communication and knowledge-oriented functions in particular user situations* (Leroyer 2007: 110). The data disposition in the ARTES dictionary takes into account different users and user situations targeted by the tool - learners of terminology and translation studies, translators, learners in scientific fields, scientists, and linguists. Translators and translation learners need to find relevant information for translating concepts and phrases which they do not necessarily completely understand. On the other hand, scientists and science students need to find information that will help them formulate their ideas in the second language. Thus two opposing situations can be distinguished, leading to the necessity to make

[11] Master professionnel ILTS (Industrie de la langue et traduction spécialisée) : http://www.eila.univ-paris-diderot.fr/formations-pro/masterpro/ilts/index

[12] http://step.ipgp.fr

[13] http://www.ipgp.fr

[14] https://artes.eila.univ-paris-diderot.fr

ARTES both an encoding and a decoding dictionary. There is yet another function of the ARTES dictionary which enables the linguists and the teachers to navigate through data by specifying criteria for data selection. This function refers to a setting where a language specialist needs to retrieve data in order to construct useful material for teaching or research purposes. As shown in the Figure 1, there are three major accesses to data: through terminology, through phraseology, and by multiple criteria query.

The function "Terminology in context" allows the user to interrogate the database for terminology which is domain-dependant and to display useful information in relation with each term. The data is categorised according to different settings: a term from the point of view of its meaning, its usage or its translation. The following function, "Discourse phraseology", provides a template for navigating through phraseology which is domain-free, yet frequently used in LSP communications, and includes collocations, collocational frameworks, expressions and other types of phraseological units (for example: *to be described elsewhere by, to be in a poor agreement with, to provide evidence for, tremendous amount of, etc.*). The access to this sort of data is provided via semantico-discursive categories which were pre-identified thorough multi-domain corpus analysis of phraseological data (Pecman 2004, 2007). The last function, "Multiple query search", intended for linguists, allows the user to retrieve data by multiple criteria and thus construct useful material for teaching or research purposes.

4. Improving ARTES database through research in terminology, phraseology and discourse analysis

The ARTES project is being developed in close relation with research in a number of connected areas: terminology, phraseology, translation, LSPs, corpus linguistics, genre and discourse analysis. The results of studies conducted in these various fields are implemented in the database whose architecture reflects the advances in our knowledge on LSPs. In this way, the solutions adopted in the ARTES database are to large extent based on switching between theories, observed linguistic evidences, and target users' needs. In the other words, we proceed by examining various theoretical premises in translation and LSP oriented terminology management and then by applying or adapting them in agreement with the observed linguistic phenomena, all to serve adequately various LSP speakers' needs defined in the context of the ARTES project. The present section presents some of the major aspects of lexical resource creation with the ARTES database for improving language-related research and applications in areas such as information retrieval, terminology, phraseology or discourse analysis.

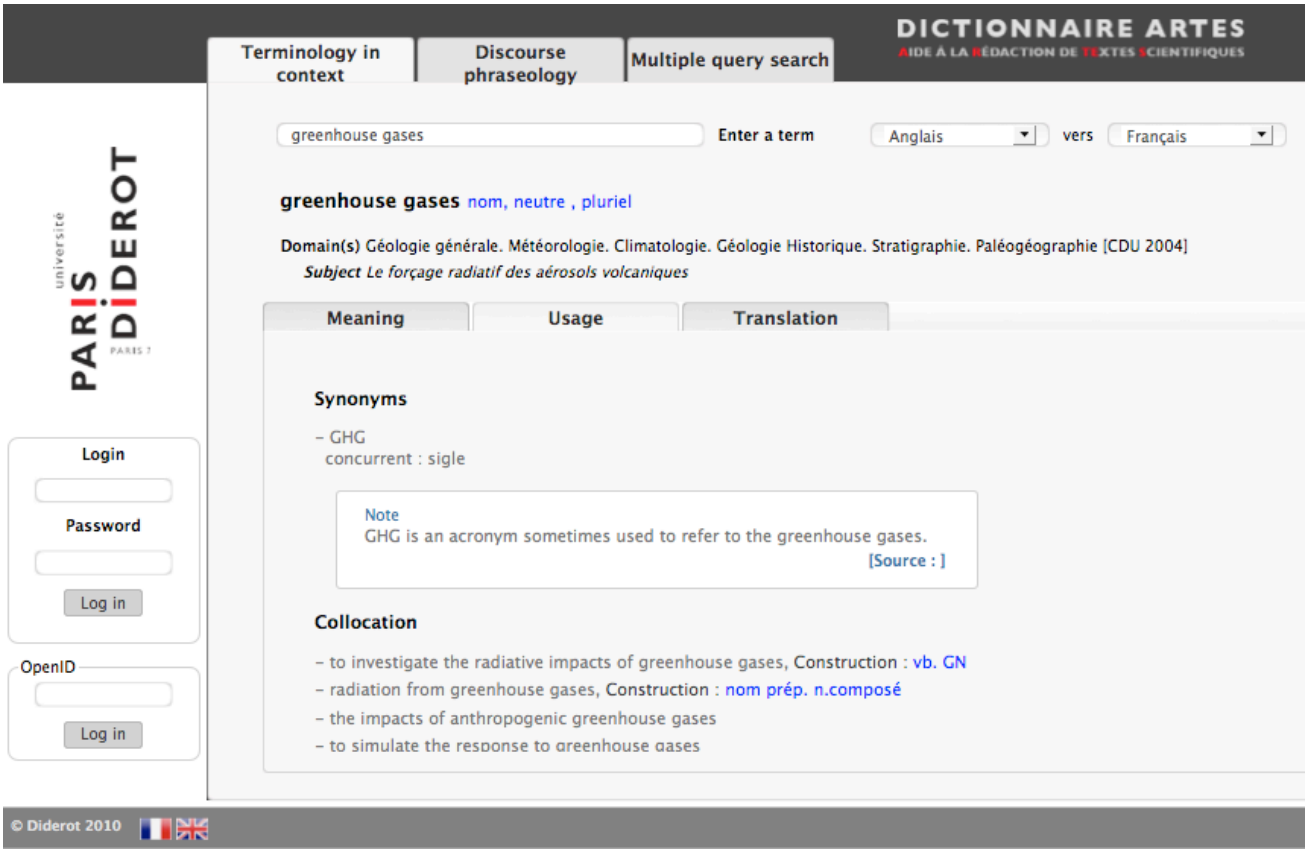


Figure 1: ARTES dictionary interface

4.1 Processing multidomain resources

Creation of lexical resources in a multidomain perspective raises the question of the organisation of lexical units according to different domains. It is well known that a term can have different meanings according to the different domains it may occur in. Assigning a term to a domain, or a series of domains, is often a complex task. Let us consider but one simple example: the word *water* which can be assigned to the domain of chemistry (where its definition could be "chemical compound consisting of two atoms of hydrogen and one atom of oxygen"), physics (where its definition could be "a liquid that changes its phase into ice at 0°C and into gas at 100°C"), or geology, climatology and meteorology (where its definition could be "the element of which seas, lakes, and rivers are composed, and which falls as rain and spouts from springs"), not to mention its role in the general language. The degree of precision of a domain specification is another difficulty we have to deal with. For example the term *fault*, defined as "a fracture in the Earth's crust that divides a geological area into two blocks which move relative to one another" can be assigned equally to the following domains: geology, seismology, plate tectonics, structural geology, geomorphology, endogenous geology, geophysics, and so on.

In the case of polysemy or in general, in an LSP database it is important to have an efficient system of descriptors for domain specifications. For the ARTES database, we have chosen to follow Universal Decimal Classification (UDC), which has systematic approach to classification, allows for exhaustiveness, and to choose a level of precision when specifying a domain. Consequently, in the ARTES database, a term can be easily assigned to one or more domains.

At this stage of the project, the resources are only being built and the domain coverage is not yet as large as the DB allows it. There are nevertheless some 23 000 terms, 25 000 collocations and 1 500 domains recorded in the DB. The collection of generic collocations is at initial stage and contains more then 100 entries.

4.2 Processing multilingual resources

So as not to be limited to a fixed number of languages when creating resources in LSPs, the architecture of the ARTES database was developed to allow for a multilingual approach. Each term can be assigned to one language specification. The table of language specifications contains some fifty languages. The pairs of equivalences can be established among any two terms or collocations of equivalent terms. For example, if *greenhouse gases* and *gaz à effet de serre* are defined as equivalents, it is then possible to align their respective collocations, e.g. *man-made green house gases* and *gaz à effet de serre anthropique, to reduce green house gases* and *réduire les gaz à effet de serre,* etc. It is thus possible to consider different types of units when working on the transfer of meaning from one language to another.

Nevertheless, establishing translational units is a very problematic matter, even in the exact sciences. In many cases the equivalences between terms are partial. We have thus added in the database a field for translation notes in order to indicate the contexts in which the equivalence is acceptable. For example, the concept of "rocks fabric" or "the fabric of a rock", in the domain of geology and mineralogy, is difficult to translate into French as it comprises the idea of rock's texture, composition and the disposition of its crystals. The French langue uses finally a loanword "fabrique" but which in common language is a false cognate meaning "factory". This difficulty is nevertheless particularly apparent in the domains where cultural differences are important. As the ARTES database is a multidomain language resource, the domains such as law, education or social sciences are also included. For example, in the domain of bankruptcy law the *distressed company* seems to be a suitable equivalent for *entreprise en difficulté*, but the cultural differences of law systems in English and French speaking countries, raise problems of translation, despite a European tendency for harmonisation, e.g. in US *distressed companies* are the matter of *bankruptcy courts* while in France *les entreprises en difficulté* are the matter of *tribunaux de commerce*. It would be though improper to say that *tribunaux de commerce* is the exact equivalent of *bankruptcy courts*.

On the other hand, when we have a series of synonymous units in a source and target language, they can all be considered as equivalent. Establishing multiple equivalent pairs is then necessary. The following examples taken from trans-discipline phraseology: *the present section concentrates on, our concern here is with, we shall concentrate here on* can be all considered as possible translations for *dans cette partie nous abordons, nous allons maintenant aborder, nous nous intéressons ici à*. In order to facilitate the processing of multiple equivalences across languages, we are currently modifying the ARTES database architecture in order to integrate synsets which can be defined within a language before aligning them across languages.

4.3 Processing terminological variation

Handling terminological variation when creating language resources is another complex matter. This issue relates to the phenomenon of synonymy, which in the ARTES project is tackled from a broad perspective and termed "concurrence", referring to a situation of competition between terms. In many cases the synonymy between terms is partial. We have thus added a series of descriptors allowing to determine the degree or the type of "concurrence" between terms: acronym, extended version of a term, reduced version of the term, partial synonymy, and so on. For example, in geology, *Moho* is a reduced version of a term for *Mohorovicic discontinuity*, *VLP* is an acronym for *very-long period*, *ice flow* and *ice creep* could be considered as partial synonyms, or near isonyms. An additional note on concurrent pairs explains the degree of superposition of

the meaning and the usage of concurrent terms.
Some more sophisticated phenomena of terminological variation are currently under study with a view to improving the method of processing terminology in the ARTES database, namely the case of complex nominal groups which appear as new terms and give rise to relatively high degree of variation within discourse: e.g. *naturally ventilated buildings* vs. *buildings ventilated naturally, buildings ventilated by natural means, buildings ventilated by natural convection*.
The current procedure for processing terminological variation contains several stages. The first one consists in accessing the nature of the variation through corpus and discourse analysis. Variation, specifically nominal variation, in LSP is generally considered as an indicator of neology. Nevertheless, in some instances, variation can play specific rhetoric or expressive effect. The second step consists in determining, again through corpus and discourse analysis, which variant is dominant, and which variants are the alternative ways of expressing the same concept. The dominant variant is encoded in the ARTES DB as a main entry, while all the relevant variations of the entry are recorded as its concurrents. The type of relation between the main term and each variant is precised, and a note is added to provide linguistic information on the usage of each variant, for instance explain the specific nature of a variant or its context, or circumstances, of use. As the ARTES DB is a relational DB, it is possible through the user's interface to search one of variants and to access the article of the main term.

4.4 Working toward a conceptual organisation of resources

The idea behind ARTES dictionary is to bypass classical alphabetic access to data by revealing multiple relations between data, some of which are particularly useful for understanding lexicon structure.
Generic, partitive, functional, instrumental, analogical... relations can be established between terms in order to highlight the conceptual organisation of lexicon of a particular domain. Retrieving data from the ARTES database in order to display lexical resources graphically is one of the perspectives we intend to develop in the near future.
By the same token, semantic preference and prosody relations, as defined by Sinclair (1987), Louw (1993) can be established between terms determining semantically cognate terms or terms sharing the same connotation. Few authors have studied these phenomena in LSPs, among them Tribble (2000), Hunston (2007) and Louw & Chateau (2010). Semantic preference and prosody have been studied extensively by the members of ARTES team (Kübler & Pecman forthcoming, Castagnoli *et al.* forthcoming) with a view to improving even further the linguistic information encoded in the ARTES database. These phenomena can indeed help us to enhance our knowledge of lexicon structure in terms of meaning and connotation.

One of the many ambitious approaches to data offered in ARTES is also the onomasiological access to collocations which are common to a variety of scientific discourses, as a help tool for drafting scientific texts (Pecman 2007, Pecman *et al.* 2010). This discourse phraseology has enriched studies on GSL (General Scientific Language) (Pecman 2004, 2007) which looks at ready-made patterns commonly employed by researchers and experts regardless of their discipline. In ARTES we have proposed 14 main classes and some 80 sub-classes for categorizing GSL phraseology in types, according to their meaning and function in LSP discourse.
Only very recently, the more comprehensive studies of a similar type of language resources, namely academic phraseology, have been carried out (cf. Durrant and Mathews-Aydınlı 2011; Simpson-Vlach and Ellis 2010). For example, Simpson-Vlach and Ellis (2010) propose an Academic Formula List (AFL) of most frequent lexical bundles found in academic communication, which are sorted according to major discourse-pragmatic functions. Nevertheless, if we compare the AFL with the collocations used in GSL, we find that the formulas used in academic setting are significantly different from those used in scientific setting. Moreover the methodologies for processing collocational phenomena for creating reusable lexical resources are still underexplored. In much the same way, the studies and resources on expert, rather then learner, trans-discipline phraseology are still lacking.

4.5 Integration of complex lexical items such as collocations, collocational frameworks and prefabricated sentence builders

In line with advances in corpus linguistics, the ARTES resources are constructed giving priority to context for determining the meaning and the usage of terminological units. Terms are considered as main entries in the database, while collocations, collocational frameworks and ready-made sentence builders are handled as secondary entries. They behave as preferential contexts of use, which provide useful information on the combination profile of terms in LSP communicative situations.
Studies on collocations and translational problems from a corpus perspective (Kübler 2003, Pecman 2004, Volanschi 2008) have encouraged us to separate specific collocations (associated with terminology) from generic collocations (associated with discourse functions). The information on specific collocations avoids collocational blends when using highly scientific or technical terms in second language communication (e.g. the adjectival term *buoyant* used in a comparative form *to be more buoyant* corresponds in French to a nominal term modified by an adjective: *avoir une plus grande flottabilité*). Similarly, the information on generic collocations allows the user to go further in achieving native-like communicative skills. The generic collocations are often associated to lexical units which are domain non-specific (e.g. *aspect,*

approach, method, study, result, limit, question, problem, evaluate, describe, etc.) with which they enter into collocation (e.g. *to raise a question*, *promising results, experimental approach, etc*.), or they act as sentence builders (e.g. *the most complete account of this problem is found in..., our conclusions focus on aspects such as...*). Both collocations, specific and generic, are analysed in the ARTES database according to their syntactic structure (e.g. *to raise a question*: vb_noun, *experimental approach*: adj_noun) and offer a very useful information for LSP users, particularly when communicating in a second language or working in translation perspective.

5. Conclusion

The ARTES database is an innovative approach to creating lexical resources where database development and an in-depth linguistic analysis of language phenomena are closely interwoven. The originality of this tool lies in its comprehensive approach to language items of relevance in LSP translation and communication, encompassing terminological, phraseological and discoursal elements. The contrastive approach to languages and to scientific disciplines extends further the coverage of lexical resources stored in the ARTES database. This multiple approach makes of the ARTES database an interesting framework for conducting research on a variety of linguistic phenomena observable in relation with LSP. Furthermore, a growing variety of LSP users (translators, teachers, students, experts and linguists) motivated the design of applications that ensure entering and retrieving information from the database by taking into account different user situations. Although designed for a dictionary type use, the ARTES database offers many possibilities for extracting lexical resources and thus anticipate new situations of use: for instance linking the terminological and phraseological data stored in the ARTES database to an online concordancer would allow to display lexical items in larger contexts for distribution analysis. The future research will focus on exploring this new orientation of research. Finally, we think that research in terminology and phraseology from a lexical resources creation perspective can lead the way to a better understanding of language in terms of cognition, description and teaching.

6. Acknowledgements

The design of the ARTES database was supported by the funds provided by the Department for Applied Linguistics and Intercultural studies - UFR EILA – at Paris Diderot University. We would like to thank Pascal Cabaud, from the team System within EILA department, for his ongoing participation in the project.

7. References

Castagnoli, S., Ciobanu, D. Kunz, K., Kübler, N., Volanschi, A. (forthcoming). Designing a Learner Translator Corpus for Training Purposes. in Kübler N. (Ed.), *Practical approaches of theorical models for language corpora and language-related teaching.* Peter Lang: Bern.

Durrant, Ph., Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1), pp. 58--72.

Froeliger, N. (2008) Le facteur local comme levier d'une traductologie pragmatique. *Meta, le Journal des traducteurs*, 55(4).

Humbley, J. (2008). Les dictionnaires de néologismes, leur évolution depuis 1945 : une perspective européenne. In J.-F. Sablayrolles (Ed.), *Néologie et terminologie dans les dictionnaires*, Paris : Honoré Champion. Collection Lexica, Mots et dictionnaires, pp. 37--60.

Hunston, S. (2007). Semantic prosody revisited. International *Journal of Corpus Linguistics*, 12(2), pp. 249-268.

Kübler, N. (2011). Working with different corpora in translation teaching. In A. Frankenberg-Garcia, L. Flowerdew, G. Aston (Eds.) *New Trends in Corpora and Language Learning.* London: Continuum.

Kübler, N. (2003). Corpora and LSP translation. In F. Zanettin, S. Bernardini & D. Stewart (Eds.), *Corpora in Translator Education*. Manchester: St Jerome Publishing, pp. 25--42.

Kübler, N. & Pecman, M. (forthcoming). The ARTES bilingual LSP dictionary: from collocation to higher order phraseology. In S. Granger & M. Paquot (Eds.) *Electronic lexicography*. Oxford : Oxford University Press.

Kübler, N.; Pecman, M., Bordet, G. (2011). La linguistique de corpus entretient-elle d'étroites relations avec la traduction pragmatique ?. In M. Van Campenhoudt, T. Lino, R. Costa (Dir.) *Passeurs de mots, passeur d'espoir: Lexicologie, terminologie et traduction face au défi de la diversité.* Actes des huitièmes journées de Lexicologie, Terminologie, traduction (LTT) 15-17 Oct. 2009 Lisbonne, 579--592.

L'Homme, M-C. (2007). De la lexicographie formelle pour la terminologie : projets terminographiques de l'Observatoire de linguistique Sens-Texte. in *Actes du colloque BDL-CA* (Bases de données lexicales : construction et applications), 23 avril 2007, OLST, Université de Montréal, pp. 29--40.

Leroyer, P. (2007). Bringing corporate dictionary design into accord with corporate image. From words to messages and back again. In Gottlieb, H. & J. E. Mogensen (Eds.) *Dictionary vision, research and practice: selected papers from the 12th International Symposium on Lexicography*, Copenhagen 2004. Terminology and Lexicography Research and Practice 10. Amsterdam/Philadelphia: John Benjamins, pp. 109--117.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). Amsterdam: John Benjamins.

Louw, B. & Chateau, C. (2010). Semantic Prosody for

the 21st Century: Are Prosodies Smoothed in Academic Context? A Contextual Prosodic Theoretical Perspective. In S. Bolasco, I. Chiari, L. Giuliano (Eds): *Statistical Analysis of Textual data: Proceedings of the tenth JADT Conference*

Pecman, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique*. Thèse de Doctorat en linguistique. Université de Nice-Sophia Antipolis.

Pecman, M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue française de linguistique appliquée.* « Lexique des écrits scientifiques », 12(2), pp. 79--96.

Pecman, M. (2008). Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions. In S. Granger & F. Meunier (Eds.) *Phraseology in language learning and teaching*. Amsterdam/Philadelphia: John Benjamins, pp. 203 --222.

Pecman, M., Juilliard, C., Kübler, N., Volanschi, A. (2010). Processing collocations in a terminological database based on a cross-disciplinary study of scientific texts. *Cahiers du Cental.* Proceedings of eLex2009, Université catholique de Louvain, Louvain-la-Neuve, Belgique, pp. 249--262.

Simpson-Vlach, R & Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistic*, 31(4), pp. 487 --512.

Sinclair, J. (1987). *Looking up: An account of the* COBUILD *project in lexical computing and the development of the Collins COBUILD English language dictionary.* London/Glasgow: Collins.

Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery, (Eds.), *Rethinking language pedagogy from a corpus perspective:* Papers from the Third International Conference on Teaching and Language Corpora. New York: Peter Lang, pp. 74--90.

Volanschi, A. (2008). Étude et modélisation des phénomènes collocationnels : Implémentation dans un système d'aide à la rédaction en anglais scientifique, Thèse de doctorat en linguistique. Université Paris Diderot.

Volanschi, A. & Kübler, N. (2011). The impact of metaphorical framing on term creation in biology. *Terminology* 17(2).

Enriching Morphological Lexica through Unsupervised Derivational Rule Acquisition

Géraldine Walther[1], **Lionel Nicolas**[2]

1. Univ. Paris Diderot, Sorbonne Paris Cité & CNRS, LLF, UMR 7110 & INRIA, Alpage, UMR-I 001
175 rue du Chevaleret, 75013 Paris, France
2. Equipe RL, Lab. I3S, Université Nice Sophia-Antipolis & CNRS
2000 route des Lucioles, BP 121, 06903 Sophia Antipolis, France
geraldine.walther@linguist.jussieu.fr, lnicolas@i3s.unice.fr

Abstract

In a morphological lexicon, each entry combines a lemma with a specific inflection class, often defined by a set of inflection rules. Therefore, such lexica usually give a satisfying account of inflectional operations. Derivational information, however, is usually badly covered. In this paper we introduce a novel approach for enriching morphological lexica with derivational links between entries and with new entries derived from existing ones and attested in large-scale corpora, without relying on prior knowledge of possible derivational processes. To achieve this goal, we adapt the unsupervised morphological rule acquisition tool MorphAcq (Nicolas et al., 2010) in a way allowing it to take into account an existing morphological lexicon developed in the Alexina framework (Sagot, 2010), such as the L*efff* for French and the L*eff*e for Spanish. We apply this tool on large corpora, thus uncovering morphological rules that model derivational operations in these two lexica. We use these rules for generating derivation links between existing entries, as well as for deriving new entries from existing ones and adding those which are best attested in a large corpus. In addition to lexicon development and NLP applications that benefit from rich lexical data, such derivational information will be particularly valuable to linguists who rely on vast amounts of data to describe and analyse these specific morphological phenomena.

1 Introduction

Among existing lexical resources, morphological resources accounting for an language's inflectional properties are very common. Resources specifying derivation phenomena and derivation links between individual lexical entries, however, appear to be less complete — even for major languages such as French and Spanish. This is not a surprising fact, since, if we look at descriptive grammars, we also notice that the potentially missing parts of a language's morphological description usually concerns derivation, while inflection is thoroughly documented.

In this paper, we use an unsupervised morphological rule acquisition tool to uncover derivation rules for French and Spanish and acquire new lexical information, namely derivation links between existing lexical entries as well as new derived lexical entries, that is missing in two of the major lexical resources existing for these two languages: the L*efff* (Sagot, 2010), a large-scale morphosyntactic lexicon for French, and the L*eff*e (Molinero et al., 2009), a large-scale morphological lexicon for Spanish. In order to uncover these derivation rules missing in these two lexica, we adapt the unsupervised morphological rule learning technique MorphAcq (Nicolas et al., 2010) enabling it to take into account lexical data and complete the set of derivation rules in the L*efff* and the L*eff*e.

In the following sections, we will first sketch an overview of existing (semi-) automatic morphological rule acquisition techniques and lexical data acquisition techniques (section 2). In section 3, we describe the lexical framework Alexina (Sagot, 2010) and the Alexina lexica we used in our experiments. Then, in section 4, we describe morphological rule acquisition using MorphAcq, the acquisition tool itself, its adaptation to account for lexical data, the input corpora and the obtained raw results. In section 5, we show that using morphological rule acquisition techniques helps enriching existing lexical resources. We finally conclude in section 6.

2 Related Work

Unsupervised methods for morphological rule acquisition can be divided into roughly two types: those that aim at building morphological analysers through the optimisation of a specific set of metrics, and those that concentrate on the explicit uncovering of morphological information.

Among the first type, the most cited are *Linguistica* (Goldsmith, 2001; Goldsmith, 2006) and *Morfessor* (Creutz and Lagus, 2005). *Linguistica* constitutes the first real attempt to use the concept of *MDL* (*Minimum Description Length*) for encoding a complete corpus w.r.t. morphemes using as few bits as possible, thus trying to achieve the best possible affix and stem recognition. In (Creutz and Lagus, 2005), the authors also use the MDL approach without restricting the analysis of a word into only one facultative prefix, only one stem and only one suffix as is the case in (Goldsmith, 2001). Morfessor has later been extended for treating allomorphisms (Kohonen et al., 2009). Later, in (Golenia et al., 2009), MDL is used to pre-select possible stems for given forms; the stems are separated from the rest and the remaining strings considered possible affixes. These possible affixes are then first broken into substrings and then re-assembled according to a metric relying on the number of these substrings' occurrences. Spiegler *et al.* (2010), Bernhard (2008) and Keshava (2006) describe methods inspired by the work of Harris (1955) and extensions thereof (Hafer and Weiss, 1974; Déjean, 1998). These approaches focus on *transition probabilities* and *letter successor variety*, i.e., the distribution of letters following a given sequence

of characters. They detect morpheme boundaries using entropy measures. The method described in (Demberg, 2007) also follows the algorithm in (Keshava, 2006), but corrects important drawbacks, in particular by handling with the fact that, for languages such as English, numerous forms are characterised by the absence of any kind of suffix. Dasgupta and Ng (2007) further extend the (Keshava, 2006) methods to the treatment of multiple suffixes.

The second type of unsupervised morphological rule acquistion methods concerns ways to identify morphological information *per se*. Thus, Lavallée and Langlais (2010) succeed in identifying word-formation using analogical processes such as *live vs. lively* and *cordial vs. cordially*. In this approach, every analogical process is weighted according to its productivity, i.e. the number of attested forms w.r.t. to the potential applicability of the analogical process. A similar approach is described in (Lignos et al., 2009). In this latter approach, however, productivity is measured according to the number of shared stems and the length of the attached affixes for each given form pair. In (Bernhard, 2010), the similarity of two forms is measured either by an edit distance or, when it is too small, by automatically extracted morphological and analogical rules. This similarity measure is then used in a *clustering* algorithm used to group possible forms for a given lemma. In (Can and Manandhar, 2009), the authors start with grouping forms according to similarity and then try to identify analogical processes between the forms of distinct groups. The productivity of the analogical processes is measured according to the number of shared stems. Finally, in (Monson et al., 2008) the morphological affixation rules applying to a given position class (in the sense of (Stump, 2001)) are directly identified, without prior identification of concrete possible affixes. This task uses a series of heuristics that control the output of the morphological rule detection method.

Concerning the acquisition of lexical data, several algorithms have been designed to extract new lemmas from a limited amount of information. They have been applied to several languages such as Russian (Oliver et al., 2003), French verbs (Clément et al., 2004), German nouns (Perera and Witte, 2005), Slovak (Sagot, 2005), Italian (Zanchetta and Baroni, 2005), French verbs, nouns and adjectives (Forsberg et al., 2006) and Polish (Sagot, 2007). These techniques differ from one another in various aspects, such as the soundness of the underlying probabilistic model and/or heuristics, the completeness of the manually described linguistic information that are exploited (e.g.,constraints on possible stems for each inflectional class, derivation patterns, etc.), the use of Google for checking the "existence" of a form, or the use of (probabilized since uncertain) part-of-speech information when it becomes available.

The acquisition of derivational links and derived lexical entries has also been studied. Systems like GéDériF (Dal and Namer, 2000) and its successors Walim (Namer, 2003) and Webaffix (Hathout, 2002) are for instance able to acquire new derived lemmas whenever their base lemma and their derivation rules are known.

In this work, we focus on the acquisition of new derived lemmas and derivation links in cases where the derivation rules have yet to be found. We propose an approach using the uncovering of these derivation rules through unsupervised morphological rule acquisition.

3 Presentation of the Input Lexica

3.1 The Alexina Framework

In our experiments, we used our morphological rule learning tool MorphAcq (described in section 4.1) jointly with lexical resources developed within the Alexina framework (Sagot, 2010). The lexica developed within the Alexina framework have the advantage of being all freely available[1] for quite a reasonable number of morphologically relatively diverse languages.

In this section, we thus briefly describe the Alexina framework underlying the two lexica we conducted our experiments on.

Although the Alexina framework covers both the morphological and the syntactic level, we only exploit the morphological level of the developed resources. Alexina allows for representing lexical information in a complete, efficient and readable way, that is meant to be independent of the language and of any grammatical formalism. It is compatible with the LMF standard[2] (Francopoulo et al., 2006). Numerous resources are being developed within this framework, such as the L*efff*, a large-coverage morphological and syntactic lexicon for French (Sagot, 2010), the L*eff*e for Spanish (Molinero et al., 2009), and also the L*eff*ga for Galician, PolLex for Polish (Sagot, 2007), SkLex for Slovak (Sagot, 2005), PerLex for Persian (Sagot and Walther, 2010), SoraLex for Sorani Kurdish (Walther and Sagot, 2010) and KurLex for Kurmanji Kurdish (Walther et al., 2010).

The Alexina model is based on a two-level representation that separates the description of a lexicon from its use:

- The intensional lexicon factorises the lexical information by associating each lemma with a morphological class (defined in a formalised morphological description) and deep syntactic information; it is used for lexical resource development;
- The extensional lexicon, which is generated automatically by *compiling* the intensional lexicon, associates each inflected form with a detailed structure that represents all its morphological and syntactic information; it is directly used by NLP tools such as parsers.

3.2 The Input Lexica

The L*efff* is the first lexical resource developed within the Alexina formalism (Clément et al., 2004) and has been continuously manually and automatically completed since then. The L*eff*e (2009) is more recent. Still, the L*eff*e contains a complete morphological description.[3]

In Alexina lexica, morphological information is encoded in a separate morphological description file that encodes the

[1] Under LGPL-LR licences, downloadable at the following address: `http://alexina.gforge.inria.fr`.

[2] Lexical Markup Framework, the ISO/TC37 standard for NLP lexica.

[3] The difference in scale between the L*efff* and the L*eff*e mainly lies in the syntactic level.

operations necessary to create the different forms for each given lemma according to a specific inflection table it belongs to. An example of inflection rules (`<form .../>`) and derivation rules (`<derivation .../>`) is given below: the inflection rule adds the suffix *es* to the stem of verbs in *–er*, indicated by the name of the table the rule belongs to. It thus creates an inflected form with the morphological tag `PS2s` (present indicative or subjunctive, second person singular). The derivation rule indicates that a derived lemma can be created from a nominal base in *-ion* by adding *ner* to the base stem.[4]

```
<table name="v-er" canonical_tag="W"
 rads="...*">
<form suffix="es" tag="PS2s"/>

<derivation suffix="ner" table="v-er"/>
```

In Alexina lexica, the relevant inflection class is specified for each lemma in the column immediately following the citation form. The lemmas are listed in a POS specific file containing the intensional lexical entries. Lemmas and inflection tables from the Le*fff*s verbal entries are represented as below[5].

```
agacer v-er:std
agir   v-ir2
```

Adding new derivation rules requires encoding the rule in the Alexina language. Adding new derived lemmas hence entails indicating their newly associated inflection table.

4 Morphological Rule Acquisition from Raw Corpora

4.1 The MorphAcq System

MorphAcq (Nicolas et al., 2010) is a tool that takes as an input raw corpus data in a given language, that is supposed concatenative,[6] and automatically computes a data-representative description of the language's morphology.
Eventhough MorphAcq is still in a preliminary state of development, it has already proven its ability to compete with the state of the art, in particular by its first participation to the MorphoChallenge (Kurimo et al., 2009) competition.
MorphAcq can be thought of as a set of filters that sequentially refines a list of (candidate) affixes and a list of sets of related affixes, which are meant to belong to the same inflectional or derivational paradigm: such sets are called *morphological families*. The combination of an affix from a morphological family and a stem associated with this morphological family is expressed as a *morphological rule*. For MorphAcq, a morphological rule, be it derivational or inflectional, consists in adding one (possibly empty) affix (prefix or suffix) to a given stem with no character deletion or substitution whithin the stem or derivational base. Linguistic phenomena that might modify the stem and/or the affix thus lead to various different morphological rules.[7]
The overall MorphAcq algorithm can be decomposed into five steps:

1. Generate an over-covering and "naive" list of candidate affixes, i.e., substrings that may be affixes. In other words, each form found in the corpus is split into a large number of stem+affix combinations (among which most are incorrect).

2. Detect candidate affix pairs that seem to be related (see discussion of step 2 below for details). For example, if affixes a, b and c belong to the same morphological family (e.g., to the same inflection class), then this step should detect pairs $\{a, b\}$, $\{b, c\}$ and $\{a, c\}$.

3. Build morphological families according to sets of pairs that share a common stem. For instance, if affixes a, b and c have all been seen on the same stem, and if the pairs $\{a, b\}$, $\{b, c\}$ and $\{a, c\}$ have been detected as "related" in step 2, the morphological family $\{a, b, c\}$ is built.

4. Split compound affixes. For example, split the English suffixes *-ingly* into *-ing* and *-ly*.

5. Detect which substrings can connect stems and split compound stems. For instance, detect that the hyphen ("-") can connect English stems and split the form "brother-in-law" into "brother + in + law".

All these steps are based mostly on simple computations with no or few free parameters. Therefore, MorphAcq can be used on virtually any concatenative language with almost no expert work.
We focus here on steps 2 and 3, which needed adaptation for this work in order to take into account external lexical data. The first step was left unchanged, and the fourth and fifth steps provide data that is not relevant here.
Step 2 exploits the following crucial observation about form- vs. lemma frequency: the frequency of a lemma's inflected forms tends to vary consistently with the the lemma's overall frequency. For example, in general texts, all inflected forms of the lemma *to talk* are more frequent than their corresponding forms from the lemma *to orate*. Moreover, this observation is not limited to the inflected forms of a lemma, but applies also derived lemmas and forms. For example, let us consider a set of forms found in the input corpus and that can be split into a stem and one of the two affixes a_1 or a_2. The goal of step 2 is to decide whether a_1 and a_2 belong to the same morphological family, i.e., whether they belong to the same inflectional or

[4] Examples are from the Le*fff*s morphological description.

[5] Syntactic information, including detailed valency information, is included in the Le*fff*, but is not shown here out of clarity reasons, as it is not relevant in this paper.

[6] We define here a concatenative language as a language that uses morphological operations that can all be entirely described through affixation. The rules are applied to graphemic sequences. Sandhi phenomena are treated independently, e.g., through the operation `<fusion .../>` in an Alexina lexicon.

[7] For instance, in French, *chantons* and *mangeons*, inflected forms corresponding to stems *chant-* and *mang-* correspond to two different morphological rules, one that adds the suffix *-ons* and another one involving the suffix *-eons*. The fact that the "real" suffix is *-ons* in both cases and that the extra *-e-* is the consequence of a phonographemic rule is not extracted.

derivational paradigm. If this is the case, which means that the frequency of lemmas and the frequency of their forms are found to vary accordingly, sorting stems s according to the frequency of the forms $s + a_1$ or according to the frequency of the forms $s + a_2$ should lead to similar orderings. Oppositely, if a_1 and a_2 are not related, both orderings should be very different.

Once pairs of related affixes are identified, step 3 builds sets of affixes that constitute morphological families by putting together pairs that have been seen on at least one common stem. It then uses four different heuristic filters for removing incorrect affix sets. Among these filters, the main one relies on the same observation as step 2. Indeed, this form- and lemma-level frequency consistency implies that the more frequent a lemma is, the more of its inflected forms will occur in the corpus. Therefore, less frequent lemmas should be attested in the corpus only by some of the inflected forms generated by their inflection class, whereas more frequent lemmas from the same inflection class are attested by more distinct forms. This means that we should be able to relate a morphological family involving n affixes with morphological families involving only $n - 1$, $n - 2$,..., 1 of these affixes, and that these families should be associated with stems of decreasing frequency. Therefore, we use a filter that keeps a morphological family with n affixes only if it at least one of its morphological subfamilies involving $n - 1$ of its affixes is identified as such.

4.2 Adapting MorphAcq

In order for MorphAcq to take into account lexical data, we modified steps 2 and 3 as follows.

First, step 2 uses the lexicon for grouping inflected forms of a same lemma, considered as a combination stem+inflection class. Instead of applying the frequency-based observation described above on two stem+affix sequence pairs, which allows to compare the two corresponding affixes, we now apply this observation on a stem+inflection class sequence and a stem+affix sequence such that the form stem+affix is not generated by the inflection class. Thus, we are able to identify affixes that are "related" to inflection classes, by means of stems they are both associated with (by the lexicon as far as the inflection class is concerned, and by the corpus as far as the affix is concerned).

For each inflection class c_b, step 3 then tries to group into affix sets the "related" affixes found during step 2. These "related" affixes generate forms that do not belong to the known inflectional paradigm of the (base) lemma l_b corresponding to their stems. They might therefore correspond to missing inflectional rules or to missing derivational rules.

We first suppose that all these rules are derivational, i.e., these forms are candidates for being inflected forms of lemmas l_d (with inflection class c_d) that are derived from that base lemma l_b. If, for at least one stem s, one of these candidate derived forms is known to the lexicon, then the lexicon provides us with its lemma l_d and inflection class c_d. This allows for computing a morphological (derivation) rule that transforms l_b into l_d. By removing the longest substring l_b and l_d have in common, we can turn this morphological rule into a generic rule that might apply to any lemma with inflection class c_b.

If this process fails on a given affix, this affix is considered inflectional: we then build the corresponding missing inflection rule. The fact that it is missing explains why the form is unknown to the lexicon although its lemma l_b is known.

Finally, MorphAcq is able to associate a confidence score with each morphological rule it outputs, based on paradigm coverage and form frequency.

	LANGUAGE	CORPUS SIZE (IN TOKENS)
Le*fff*	*French*	*∼18 215 000*
Le*ffe*	*Spanish*	*∼540 000*

Table 1: Corpora used as an input to MorphAcq

4.3 The Input Corpora

As input data to MorphAcq, we used a corpus extracted from the French newspaper *le Monde diplomatique* [8] for French, and the raw data of the *Ancora* corpus (Taulé et al., 2008) for Spanish. We were able to detect several missing derivational rules for both our input lexica. The corresponding figures are given in Table 1.

4.4 Results and Evaluation of the Morphological Rule Acquisition

When we first confronted the output of MorphAcq with the forms generated with the two Alexina lexica, the results showed that both resources seem to reasonably well encode the inflectional system of both languages. The inflectional rules that were suggested as missing rules were the result of isolated typographical errors or English loanwords. Therefore, we simply ignored the few inflection rules that were suggested by MorphAcq.

MorphAcq generated 3,131 derivational rules from our Spanish data, and 36,430 derivational rules from our French data. This huge difference is mostly due to the fact that the French corpus we gave as an input to MorphAcq is over 30 times bigger than the Spanish one. However, many of these rules have to be considered as noise. This is why we applied various filters before using them in practical lexicon enrichment experiments, as explained in the next section.

5 Enriching Lexical Resources through Automatic Acquisition of Morphological Rules

5.1 Evaluation of Acquired Derivation Rules through External Information

Derivation is a morphological process that generates a new lemma from the derivation-base of a first one. The new lemmas are part of the set of lexical entries available in the lexicon of a given language. They have to be associated with the right inflection tables since they are themselves possibly inflectable. Recall that in Alexina

[8] `http://www.monde-diplomatique.fr/`, February 2011.

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
basculement	*nc-2m*	basculer	v-er:std
centreur	*nc-2m*	centrer	v-er:std
crochetage	*nc-2m*	crocheter	v-er:std
déloyalement	*adv*	déloyal	adj-al4
fasciste	*nc-2*	fasciser	v-er:std
gourmand	*nc-2f*	gourmander	v-er:std
insolation	*nc-2f*	insoler	v-er:std
minimaliser	*v-er:std*	minimal	adj-al4
perfectionnement	*nc-2m*	perfectionner	v-er:std
reboisement	*nc-2m*	reboiser	v-er:std
soûler	*v-er:std*	soûl	adj-4
trébuchement	*nc-2m*	trébucher	v-er:std

Table 2: Examples of French derivation links acquired automatically

lexica, the inflection class is specified for each lemma in the column immediately following the citation form (the above example is simplified, since the syntactic — e.g., valency — information is not shown).

```
agacer v-er:std
agir   v-ir2
```

Before adding derivation rules to the morphological descriptions underlying the L*efff* and the L*eff*e, we first filtered out from the derivation rules output by MorphAcq those that seemed less likely, in the following way. First, we automatically filtered the output given by MorphAcq using a beam filter: for a given morphological family (including the associated base inflection class), many derivation rules may be suggested by MorphAcq, each affix in the morphological family being covered by more than one of these derivation rules (each derivation rule, in turn, usually covers more than one affix, as it creates a derived lemma that has several inflected forms). For each affix in the considered morphological family, we identify the suggested morphological rule that has the best score among those that cover that affix: it is the affix's best rule. Then, we only keep those morphological rules that are the best rules for at least one of its affixes.

Among the remaining derivation rules, we require that suffixation rules be suggested for at least two distinct morphological families and prefixation rules by 25 morphological families for French and five for Spanish.[9]

Then we automatically added all remaining derivation rules into the L*efff*'s or the L*eff*e's morphological description.

We were also able to retrieve the possible *variant* a new lemma belongs to: variants are used in Alexina to differentiate lemmas that show particular morphotactic properties with minor impact on the lemmas inflection.[10] Hence, derivation rules are represented as follows:

```
<derivation suffix="ner" table="v-er"
 var="std" />
```

Converting and filtering MorphAcq's output led to the introduction of 823 derivation rules into the French morphological description. These new rules are scattered over most existing inflection classes. For Spanish, only 132 derivation rules could be identified and added. This difference in scale has again to be imputed to the difference in size the the corpora used as input to MorphAcq.

Once the new derivation rules added into the lexica, we generated all possible derived lexical entries by applying to each existing entry all derivation rules associated with its inflection class. We obtained as large a result as 2.9 million candidate entries for French and 1.0 million candidate entries for Spanish. However, Alexina inflection tables are often associated with constraints on stems: e.g., French adjectives inflecting according to class `adj-n4` in the L*eff f*, such as *parisien(s) / parisienne(s)*, are requested to have a stem ending in *n*. Trying to inflect the new derived lemmas hence allowed us to discard all those new lemmas whose stem was not compatible with the inflection class suggested by MorphAcq.

The remaining derived lemmas were used in two different ways. First, derived lemmas that correspond to existing entries in the L*efff* or the L*eff*e were preliminarily validated as correct derived lemmas, i.e., we considered that derivation links between base and derived lemmas could be added. The entries corresponding to derived lemmas thus received a derivation link of the form *derived from X*.[11] At this point 16,646 derivational link candidates were added for French and 10,745 for Spanish.

Among the candidates, the derived lemmas are necessarily correct as lexical entries, since they were found within the lexica. Only the correctness of the derivation links with the base lemma needs to be assessed. To do so, we performed manual evaluation on randomly selected samples containing 100 candidates. For Spanish, all 100 morphosemantic links were correct (see Table 3). For French, we obtained 92 correct links out of 100 (see Table 2), but from a larger set of candidates (errors are shown in Table 4). It also became apparent that the longer the base and/or the derived lemma is, the greater the certainty of the established link's correctess. Indeed, Table 4 shows that most errors involve at least one relatively short lemma.

5.2 Using Newly Acquired Rules for Enriching Large Scale Resources

Once the derrivation links between the lemmas already contained within the L*efff* and the L*eff*e had been identified, we developed a procedure for adding new (unknown) derived lemmas (with their corresponding derivational links that initially led to suggesting them). For selecting which derived lemmas had to be added, we used form frequency information extracted from large-scale corpora.

[9] The apparent striking difference in the selectivity imposed on prefixation rules between French and Spanish comes from the different scales of the acquisition corpora fed into MorphAcq. Using the same threshold for both languages would have led to either to much noise in the French data or to few acquirable rules for Spanish.

[10] See for instance French verbs that double their stems last consonants when preceding certain suffixes: infinitive *je***t***er* "throw" vs. P1sg of the present indicative *je je***tt***e* "I throw".

[11] This tag is meant to facilitate future use of the L*efff* as a resource for studies on derivational relations.

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
calcular	*V2*	calculadamente	R1
conspirador	*N8*	conspirar	V2
desencadenante	*N1*	desencadenar	V2
extremo	*N1*	extremar	V2
horadable	*A2*	horadar	V2
justo	*N4*	justar	V2
modoso	*A1*	modosamente	R1
patrimonialista	*A2*	patrimonial	A3
racional	*A3*	ración	N3
rotulista	*N6*	rotular	A3
temperado	*A1*	temperadamente	R1
zanqueador	*N8*	zanquear	V2

Table 3: Examples of correct Spanish derivation links acquired automatically

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
attiser	*v-er:std*	attiquement	adv
bafouiller	*v-er:std*	bafouer	v-er:std
cotte	*nc-2*	coter	v-er:std
entassement	*nc-2m*	enter	v-er:std
must	*nc-2m*	muser	v-er:std
présentement	*adv*	présenter	v-er:std
salement	*adv*	saler	v-er:std
sire	*nc-2m*	sirex	nc-1m

Table 4: Examples of incorrect French derivation links. Most links involve at least a "short" lemma

For French, we used a part of the *Est Républicain* corpus[12], composed of newspaper articles published in 1999. We tokenized the corpus of the *Est Républicain* into 37.5 million tokens using the "light" version of the shallow processing chain SXPipe which is included in the distribution of the MElt POS-tagger (Denis and Sagot, 2009). For Spanish, we used a cleansed dump of the Spanish Wikipedia[13]. The Spanish Wikipedia was tokenized with the same "light" version of SXPipe. We retained the first 10 million tokens.

We used these corpora as follows. First, we filtered out candidate derived lemmas whose canonical form is not attested in the corpus. This first filtering reduced the number of derivation candidates from respectively 2.9 million and 1.0 million derived lemma candidates to 62,158 for French and 22,814 for Spanish. Then, we inflected all these candidates, generating 191,000 possible new inflected entries for French and 94,000 for Spanish. We associated those with two basic sources of information: whether each inflected form is known to the lexicon or not, and its number of occurrences, if any, in the corpus.

Then, we ranked the remaining candidates in the following iterative way: at each step, we computed a score for each derived lemma candidate by adding contributions for every one of its inflected forms; these contributions were computed as their number of occurrences, taken positively if the form is unknown to the lexicon and negatively if it is known to the lexicon. The idea of this ranking is to suggest only those new lemmas that have the best coverage of corpus forms still unknown to the lexicon and do not at the same time cover forms already known to the lexicon. After having ranked all candidates, we output the best one. All its inflected forms were now considered as known to the lexicon. This means we needed to re-compute the scores and iterate the process[14]. Each iteration outputs one candidate. We stopped when the best candidate had a score smaller or equal to 1. As a result, we obtained 1,511 new derived lemmas for French and 563 new derived lemmas for Spanish. We added these new lemmas to the L*efff* and the L*effe* respectively, specifying the corresponding derivation tag for each {base lemma, derived lemma} pair.

After adding the new lemmas we performed a small manual evaluation on 100 randomly chosen new lemmas and their derivation tags for both languages. Examples of correct new derived lemmas are shown in Tables 5 and 6, whereas the quantitative results of this evaluation are given in Table 7.

[12] http://www.cnrtl.fr/corpus/estrepublicain/

[13] http://download.wikimedia.org/eswiki/latest/eswiki-latest-pages-articles.xml.bz2, dump from Feb 3, 2011.

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
maltraitance	*nc-2f*	maltraiter	v-er:std
recapitalisation	*nc-2f*	recapitaliser	v-er:std
incinérable	*adj-2*	incinérer	v-er:std
rissolette	*nc-2f*	rissoler	v-er:std
abreuvement	*nc-2m*	abreuver	v-er:std
rétractable	*adj-2*	rétracter	v-er:std
plastification	*nc-2f*	plastifier	v-er:std
tronçonnement	*nc-2m*	tronçonner	v-er:std
grenailleur	*nc-2m*	grenailler	v-er:std
désencadrement	*nc-2m*	désencadrer	v-er:std
regardable	*adj-2*	regarder	v-er:std
grêleux	*nc-x3*	grêler	v-er:std

Table 5: Examples of new French derived lemmas acquired automatically

DERIVED LEMMA	TABLE	BASE LEMMA	TABLE
orbitador	*N5*	orbitar	V2
presentacional	*A3*	presentación	N3
correlacional	*A3*	correlación	N3
insercional	*A3*	inserción	N3
confrontante	*N1*	confrontar	V2
agudismo	*N1*	agudizar	V3
multidireccionalidad	*N7*	multidireccional	A3
distintal	*N5*	distinto	A1
letalidad	*N7*	letal	A3
aleteador	*N5*	aletear	V2
aconfesionalidad	*N7*	aconfesional	A3
zanqueador	*N8*	zanquear	V2

Table 6: Examples of new Spanish derived lemmas acquired automatically

[14] Of course, we did not recompute all scores, but only updated those which had been affected by the last output.

FRENCH	• 42 correct lemmas & derivation links, • 1 correct lemma with false derivation link, • 14 correct canonical forms with incorrect inflection tables, • 10 incorrect lemmas due to the presence of English words in the corpus, • 28 incorrect lemmas due to typographical errors in the corpus, • 5 other incorrect candidates.
SPANISH	• 40 correct lemmas & derivation links, • 7 correct canonical forms with incorrect inflection tables, • 39 incorrect lemmas due to the presence of English words in the corpus, • 9 incorrect lemmas due to typographical errors in the corpus, • 5 other incorrect candidates.

Table 7: Derived Lemma Evaluation

6 Conclusion and Future Work

In this paper, we have presented a novel method for enriching large-scale lexica with concrete derivation links and a straightforward manner to use the acquired explicit derivational information to increase a lexicon's coverage. The new derivation rules have been acquired through a specifically adaptated version of the unsupervised morphological rule acquisition tool MorphAcq.

An obvious interesting side result of this method is that the lexica on which our method has been applied now show an improved quality: derivation links have been specified within the L*eff* and L*eff*e, hence allowing to use both resources for theoretical and descriptive linguistic studies on derivation.[15]

A further step in enriching lexical resources (in general, and Alexina lexica in particular) should be to combine the morphological rule acquisition tool MorphAcq with other methods designed for identifying new possible lemmas, as described in (Sagot, 2005). We plan on running the tools developed by Sagot (2005) jointly with MorphAcq. These lemma acquisition methods that rely on information from the morphological description should benefit from the improved description provided by MorphAcq's output. MorphAcq will in return benefit from being combined with resources with greater coverage. In particular the identification of the correct inflection classes for new derived lemmas should be significantly improved. Thus, using morphological rule acquisition and lemma acquisition techniques iteratively seems a promising way for efficient lexical resource enriching. This method should help rapidly developing new lexica with completely automatic methods, hence giving access to new resources for undescribed languages.

[15]The results are freely available on `http://www.linguist.univ-paris-diderot.fr/~gwalther/homepage/Publications_(en).html`.

Acknowledgements

This work was supported in part by the EDyLex project funded by the French National Research Agency (grant number ANR-09-CORD-008).

7 References

Delphine Bernhard. 2008. Simple morpheme labelling in unsupervised morpheme analysis. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the CLEF (revised selected papers)*, pages 873–880. Springer-Verlag, Berlin, Heidelberg.

Delphine Bernhard. 2010. MorphoNet: Exploring the use of community structure for unsupervised morpheme analysis. In *Multilingual Information Access Evaluation, 10th Workshop of the CLEF (revised selected papers)*, Corfu, Greece. Springer.

Burcu Can and Suresh Manandhar. 2009. Clustering morphological paradigms using syntactic categories. In *CLEF*, pages 641–648.

Lionel Clément, Benoît Sagot, and Bernard Lang. 2004. Morphology Based Automatic Acquisition of Large-coverage Lexica. In *Proceedings of LREC'04*, pages 1841–1844, Lisbon, Portugal.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. In *Helsinki University of Technology*.

Georgette Dal and Fiammetta Namer. 2000. Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations. *TAL*, 41-2:423–446.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *NAACL HLT 2007: Proceedings of the Main Conference*, pages 155–163.

Hervé Déjean. 1998. Morphemes as necessary concept for structures discovery from untagged corpora. In *NeMLaP3/CoNLL '98: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 295–298, Sydney, Australia.

Vera Demberg. 2007. A language-independent unsupervised model for morphological segmentation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 920–927, Prague, Czech Republic, June.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong-Kong, China.

Markus Forsberg, Harald Hammarström, and Aarne Ranta. 2006. Morphological lexicon extraction from raw text data. In *Proceedings of FinTAL 2006, LNAI 4139*, pages 488–499, Turku, Finland. Springer-Verlag.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Genoa, Italy.

John Goldsmith. 2001. Unsupervised learning of the

morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

John Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Nat. Lang. Eng.*, 12(4):353–371.

Bruno Golenia, Sebastian Spiegler, and Peter Flach. 2009. Ungrade: Unsupervised graph decomposition. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece*, September.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10(11-12):371–385.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Ludovic Hathout, Nabil et Tanguy. 2002. Webaffix: finding and validating morphological links on the WWW. In *Proceedings of LREC'02*, pages 1799–1804, Las Palmas de Gran Canaria, Spain.

Samarth Keshava. 2006. A simpler, intuitive approach to morpheme induction. In *In PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, pages 31–35.

Oskar Kohonen, Sami Virpioja, and Mikaela Klami. 2009. Allomorfessor: towards unsupervised morpheme analysis. In *Evaluating systems for multilingual and multimodal information access, 9th Workshop of the CLEF*, pages 975–982, Berlin, Heidelberg. Springer-Verlag.

Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of morpho challenge 2009. In *Multilingual Information Access Evaluation, 10th Workshop of the CLEF (revised selected papers)*, CLEF'09, pages 578–597, Berlin, Heidelberg. Springer-Verlag.

Jean-Francois Lavallée and Philippe Langlais. 2010. Unsupervised morphology acquisition by formal analogy. In *Lecture Notes in Computer Science*.

Constantine Lignos, Erwin Chan, Mitchell P. Marcus, and Charles Yang. 2009. A rule-based acquisition model adapted for morphological analysis. In *Evaluating systems for multilingual and multimodal information access, 9th Workshop of the CLEF*, pages 658–665.

Miguel Ángel Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The L*eff*e. In *Proceedings RANLP 2009*, Borovets, Bulgaria.

Christian Monson, Alon Lavie, Jaime Carbonell, and Lori Levin. 2008. Evaluating an agglutinative segmentation model for paramor. In *SigMorPhon '08: Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 49–58, Morristown, NJ, USA.

Fiammetta Namer. 2003. WaliM : valider les unités morphologiquement complexes par le Web. In B. Fradin et al., editor, *Silexicales 3 : les unités morphologiques*, pages 142–150, Villeneuve d'Ascq, France. Presses Universitaires du Septentrion.

Lionel Nicolas, Jacque Farré, and Miguel A. Molinero. 2010. Unsupervised learning of concatenative morphology based on frequency-related form occurrence. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, Helsinki, Finland, September.

Antoni Oliver, Irene Castellón, and Lluís Màrquez. 2003. Use of Internet for augmenting coverage in a lexical acquisition system from raw corpora: application to Russian. In *Proceedings of the RANLP'03 International Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria.

Praharshana Perera and René Witte. 2005. A self-learning context-aware lemmatizer for German. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 636–643, Vancouver, Canada.

Benoît Sagot and Géraldine Walther. 2010. A Morphological Lexicon for the Persian Language. In *Proceedings of LREC'10)*, Valetta, Malta. ELDA.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658, Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic, September. Springer-Verlag.

Benoît Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of the 3rd Language & Technology Conference*, pages 423–427, Poznań, Poland, October.

Benoît Sagot. 2010. The L*efff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of LREC'10*, Valetta, Malta.

Sebastian Spiegler, Bruno Golenia, and Peter Flach. 2010. Unsupervised word decomposition with the promodes algorithm. In *Multilingual Information Access Evaluation, Lecture Notes in Computer Science*, volume I. Springer Verlag, February.

Gregory T. Stump. 2001. *Inflectional Morphology. A Theory of Paradigm Structure*. CUP, Cambridge, UK.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of LREC'08*, Marrakesh, Morroco.

Géraldine Walther and Benoît Sagot. 2010. Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop (LREC'10 Workshop)*, Valetta, Malta.

Géraldine Walther, Benoît Sagot, and Karën Fort. 2010. Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrad, Serbia.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK. University of Birmingham.