Formalising Natural Languages with Nooj 2014

Formalising Natural Languages with Nooj 2014

Edited by

Johanna Monti,
Max Silberztein,
Mario Monteleone
and Maria Pia di Buono

Selected papers from the NooJ 2014
International Conference
University of Sassari, 3-5 June 2014

Cambridge
Scholars
Publishing

Formalising Natural Languages with Nooj 2014

Edited by Johanna Monti, Max Silberztein, Mario Monteleone
and Maria Pia di Buono

This book first published 2015

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-7558-9
ISBN (13): 978-1-4438-7558-5

TABLE OF CONTENTS

Part I: Vocabulary and Morphology

Part II: Syntax and Semantics

Part III: Applications

EDITORS' PREFACE

NooJ is a linguistic development environment that provides tools for linguists to construct linguistic resources that formalise a large gamut of linguistic phenomena: typography, orthography, lexicons for simple words, multiword units and discontinuous expressions, inflectional and derivational morphology, local, structural and transformational syntax, and semantics.

For each resource that linguists create, NooJ provides parsers that can apply it to any corpus of texts in order to extract examples or counter-examples, to annotate matching sequences, to perform statistical analyses etc. NooJ also contains generators that can produce the texts that these linguistic resources describe, as well as serving as a rich toolbox that allows linguists to construct, maintain, test, debug, accumulate and reuse linguistic resources.

For each elementary linguistic phenomenon to be described, NooJ proposes a set of computational formalisms, the power of which ranges from very efficient finite-state automata to very powerful Turing machines. This makes NooJ's approach different from most other computational linguistic tools that typically offer a unique formalism to their users.

Since its first release in 2002, NooJ has been enhanced with new features every year. Linguists, researchers in Social Sciences and more generally all professionals who analyse texts have contributed to its development and participated in the annual NooJ conference. In 2013, a new version for NooJ was released, based on the JAVA technology and available to all as an open source GPL project. Moreover, several private companies are now using NooJ to construct business applications in several domains, from Business Intelligence to Opinion Analysis.

The present volume contains 22 articles selected from the 53 papers presented at the International NooJ 2014 Conference, which was held from June 2nd to 4th at the University of Sassari in Sardinia (Italy). These articles are organised in three parts: "Vocabulary and Morphology" containing seven articles; "Syntax and Semantics" containing seven articles; "NooJ Applications" containing eight articles.

The articles in the first part involve the construction of dictionaries for simple words, multiword units as well as the development of morphological grammars:

- — Max Silberztein's article "The DEM and the LVF dictionaries for NooJ" gives a glimpse at how 'ideal' NooJ electronic dictionaries will look: he presents the recently released linguistic DEM and LVF dictionaries, and shows what work will have to be done in order to convert them into NooJ electronic dictionaries.
- — Hajer Cheikhrouhou's article "The Formalization of Movement Verbs for Automatic Translation using NooJ Platform" shows the author's effort to add an Arabic translation to the movement verbs described in the LVF dictionary and the application of the resulting bilingual electronic dictionary to machine translation.
- — Maximiliano Duran's article "Morphological and Syntactic Grammars for the Recognition of Verbal Lemmas in Quechua" presents an electronic dictionary for verbs in Quechua associated with a very powerful morphological engine.
- — Serena Pelosi and Alessandro Maisto's article "A Lexicon-Based Approach to Sentiment Analysis. The Italian Module for NooJ" presents a set of specialised dictionaries aimed at the automatic recognition of sentiments expressed in Italian texts.
- — Zoe Gavriilidou, Lena Papdopoulou and Elina Chatjipapa's article "<Material> Adjectives in Greek NooJ Module" describes the construction of a specialised electronic dictionary.
- — Matea Srebacic, Krešimir Šojat and Božo Bekavac's article "Croatian Derivational Patterns in NooJ" shows how the authors have solved the problem of linking a lexical entry to a large number of morphological forms in Croatian.
- — Mario Monteleone and Maria Pia Di Buono's article "The Inflection of Italian Pronominal Verbs" describes an elegant solution to the formalisation of the conjugation of pronominal verbs in Italian.

The articles in the second part involve the construction of syntactic and semantic grammars:

- — Annibale Elia and Alberto Maria Langella's article "Semantic Role Labelling with NooJ: Communication Predicates in Italian" shows a set of linguistic lexical and syntactic resources that can be used to automatically annotate expressions of communication in Italian texts.
- — Valerie Collec-Clerc's article "Recognition of Honorific Passive Verbal Form in Japanese with NooJ" shows a set of syntactic grammars capable of identifying Honorific Passive Forms in Japanese texts.

— Mourad Aouini's article "A NooJ Module for Named Entity Recognition in Middle French Texts" presents a set of grammars used to identify named entities in a corpus of Middle French texts.
— Azeddine Rhazi's article "Morpho-syntaxical based recognition of Arabic MWUs with NooJ" presents a set of grammars used to identify and extract Arabic multiword units from texts.
— Lena Papadopoulou's article "Local Grammars for Pragmatemes in NooJ" presents a set of syntactic grammars that recognise pragmatemes in Greek texts.
— Tatsiana Okrut, Yuras Hetsevich, Boris Lobanov and Yauheniya Yakubovich's article "Resources for Identification of Cues with Author's Text Insertions in Belarusian and Russian Electronic Texts" presents a set of linguistic resources that can be used to identify cues in Belarusian and in Russian texts.
— Simona Messina and Alberto Maria Langella's article "Paraphrases V↔N↔A in one Class of Psychological Predicates" presents a set of lexical and syntactic grammars that can be used to produce paraphrases of psychological predicates.

The articles in the third part describe various NLP applications based on the use of NooJ's linguistic engine:

— Božo Bekavac, Kristina Kocijan and Marko Tadić's article "Near Language Identification Using NooJ" shows how to automatise NooJ to automatically identify the language of a text.
— Hayet Ben Ali, Hela Fehri and Abdelmajid Ben Hamadou's article "Translating Arabic Relative Clauses into English using NooJ Platform" shows an interesting Machine Translation application for NooJ, and compares its results with the ones produced by Google Translate.
— Alena Skopinava, Yuras Hetsevich and Julia Borodina's article "Converting Quantitative Expressions with Measurement Units into an Orthographic Form, and Convenient Monitoring Methods for Belarusian" presents an automatic application for Belarusian texts that converts quantitative expressions into an orthographic form.
— Julia Frigière and Sandrine Fuentes' article "Pedagogical Use of NooJ dealing with French as a Foreign Language" shows how the authors use NooJ as a Lab tool to teach French at the Autonomous University of Barcelona.
— Kristina Kocijan and Marko Požega's article "Building Family Trees with NooJ" presents a series of complex grammars that can

automatically identify family relations of persons described in texts.

— Johanna Monti, Mario Monteleone and Maria Pia di Buono's article "A Knowledge-Based CLIR Model for Specific Domain Collections" presents the CLIR model and how it can be used to automatically collect and classify texts.

— Maria Pia di Buono and Mario Monteleone's article "Knowledge Management and Extraction from Cultural Heritage Repositories" presents an application that can mine texts on cultural heritage in order to automatically build a knowledge base of the domain.

— Slim Mesfar and Essia Bessaies' article "Automatic Document Classification and Event Extraction in Standard Arabic" shows a set of semantic grammars capable of identifying events in Arabic texts, and how this can be used to automatically classify documents.

This volume should be of interest to all users of the NooJ software because it presents the latest development of the software as well as its latest linguistic resources. To date, there are NooJ modules available for over 50 languages; more than 3,000 copies of NooJ are downloaded each year.

Linguists as well as Computational Linguists who work on Arabic, Belarusian, English, French, Greek, Italian, Japanese, Quechua, or Russian will find in advanced, up-to-the-minute linguistic studies for these languages this volume.

We think that the reader will appreciate the importance of this volume, both for the intrinsic value of each linguistic formalisation and the underlying methodology, as well as for the potential for developing NLP applications along with linguistic-based corpus processors in the Social Sciences.

The Editors

PART I:

VOCABULARY AND MORPHOLOGY

THE DEM AND LVF DICTIONARIES IN NOOJ

MAX SILBERZTEIN

Abstract

We have integrated Jean Dubois and Françoise Dubois-Charlier's DEM and LVF dictionaries into the NooJ linguistic software. We discuss their applications for Natural Language Processing applications.

Introduction

The NooJ project is aimed at constructing a large-coverage formalised description of natural languages. The project has two parts: (1) to represent the standard vocabulary of languages and (2) to describe how to combine the elements of the vocabulary in order to construct phrases, sentences and, more generally, to carry complex meaning. The vocabulary is a finite set of units called *Atomic Linguistic Units* (or ALUs) and is represented by an Electronic Dictionary.

Several Electronic Dictionaries exist for the vocabulary of French. For example, the Lexicon-Grammar of Verbs, developed at the LADL laboratory (*Laboratoire d'Automatique Documentaire et Linguistique*), describes the syntactic properties of over 12,000 verbs of standard French vocabulary. Entries that share a number of properties are grouped into tables; for instance, intransitive verbs (structure $\mathbf{N_0\ V}$) are stored in Tables 31x, direct transitive verbs (structure $\mathbf{N_0\ V\ N_1}$) are stored in Tables 4, 6 and 32x, etc.[1] Following the same model, other elements of the vocabulary have been described: there are Lexicon-Grammar tables for adjectives, adverbs, conjunctions, support-verbs and frozen expressions, and there are also Lexicon-Grammar tables for languages other than French[2].

Lexicon-Grammars are exhaustive and very precise. However, they are not autonomous linguistic resources, so automatic parsers cannot use them to perform any type of linguistic analysis. In particular, Lexicon-Grammars do not contain the minimal orthographical or morphological

[1] (Leclère 1990) presents the classification of the verbs in the Lexicon-Grammar.

[2] (Leclère 1998) lists works on various lexicon-grammars.

information[3] necessary to perform even a basic lexical analysis of texts. The LADL also developed the DELA[4] system of Electronic Dictionaries, which has been used successfully to parse large corpora of texts[5]. The DELA covers the standard vocabulary, but it only contains morphological information, more precisely, inflectional morphology[6].

The DELA system of dictionaries

The DELA has been designed to list all the elements of standard French vocabulary and describe their inflectional morphology. Its main components are:

- the DELAS dictionary which describes the inflection of simple words
- the DELAC dictionary which describes the inflection of multiword units

In these two electronic dictionaries, each lexical entry is associated with a series of codes that represent its morpho-syntactic properties. For instance, the following is a typical lexical entry from the DELAS[7] (cf. Courtois, 1990):

abaisser,V3+tr+z1

The lexical entry *abaisser* (to lower something) is a verb (code **V**) that conjugates according to paradigm **V3** (the same paradigm as *aimer*); it is a transitive verb (**+tr**) and it belongs to basic French vocabulary (**+z1**). The DELAS dictionary contains approximately 130,000 entries. Following is a lexical entry from the DELAC (cf. Silberztein 1990):

[3] Several tables of the lexicon-grammar contain verbs that are semantically similar, but they also contain verbs that have very different meanings. (Courtois, 1990) poses the problem of merging the DELAS and the lexicon-grammar; (Silberztein 1990) integrated in the DELAC dictionary a list of adverbs described in several lexicon-grammar tables, but these projects have not been pursued.

[4] Cf. (Courtois, Silberztein Ed. 1990).

[5] Cf. (Silberztein 1993).

[6] There are other electronic dictionaries that are similar to the DELA dictionary. In particular, the DM dictionary, included in NooJ, combines lexical entries from the DELAS and from the Morphalou dictionaries, cf. (Trouilleux 2012).

[7] Blandine Courtois constructed the DELAS dictionary (*Dictionnaire Electronique du LADL pour les mots Simples*), with some help from Jean Dubois.

pomme de terre,N+NDN+Conc+z1

The lexical entry *pomme de terre* (potato) is a noun (code **N)**; its structure is **NDN** (Noun *de* Noun). It represents a concrete noun (**+Conc**) and belongs to the basic French vocabulary (**+z1**). The DELAC dictionary contains over 300,000 lexical entries, most of which are compound nouns. Thanks to the description of the inflectional paradigm of each entry in the DELAS and the DELAC, the INTEX software[8] could automatically produce the list of all the corresponding forms for each entry of these two dictionaries: the DELAF contains the list of all the inflected forms that correspond to entries of the DELAS dictionary, whereas the DELACF contains the list of all the inflected forms that correspond to DELAC dictionary entries. The DELA system of electronic dictionaries still constitutes, over twenty years later, a reference among the electronic dictionaries used by NLP applications. However, it does not satisfy some of the requirements of the NooJ linguistic project.

From the DELA to NooJ

The first problem with the DELA is that it was not designed to be an autonomous linguistic resource, but rather to complement the Lexicon-Grammars. But these two databases cannot be merged, as they describe lexical entries and properties that are not comparable. For instance, in the DELAS dictionary, there are two lexical entries for the verb *voler*:

voler,V3
voler,V3U

The inflectional code **V3** is used to produce all the conjugated forms of the verb *voler*, including the four forms of the past participle *volé, volée, volés, volées* (eg in *Les fleurs que tu as volées*). The code **V3U** is used to produce only one form for the past participle: *volé* (eg *L'avion a volé au dessus de la Sibérie*). The lexical entry associated with the code **V3** corresponds to three entries of the Lexicon-Grammar:

Luc vole un cendrier à Marie. Le commerçant vole Luc de 10 euros.
Tu ne l'as pas volée !

[8] Cf. http://intex.univ-fcomte.fr. (Silberztein 1993) presents the DELA and the INTEX software. INTEX has been used as a linguistic tool as well as the linguistic engine of several Natural Linguistic Processing software applications, cf. for instance (Fairon Ed. 1999).

The lexical entry associated with the code **V3U** corresponds to two other entries of the Lexicon-Grammar:

L'avion vole vers Paris. La porte vole en éclats.

Two lexical entries in the DELAS correspond to five lexical entries in the Lexicon-Grammar. This situation is general for all levels of the linguistic description. For instance, the DELAS dictionary contains over 1,000 artefacts that are components of multiword units, which are listed independently in the DELAC dictionary (such as '*parce*' in *parce que*). Some entries of the DELAC are also listed in Lexicon-Grammar tables for adverbs as well as in the tables for conjunctions, but there is no way to know if they represent the same ALU, or if they represent different meanings ie different ALUs. There is no direct way to connect frozen expressions listed in the Lexicon-Grammar to their components listed in a DELA-type dictionary, etc.

The NooJ project requires a new type of electronic dictionary, the goal of which is to exhaustively formalise the vocabulary of the language. In order to formalise the vocabulary of a language, we need an electronic dictionary in which (1) all ALUs (simple words, multiword units and expressions) are described in a unified way, (2) there is an explicit link between all orthographical, morphological, syntactic and semantic properties for each lexical entry, and (3) there is an equivalence between ALUs and lexical entries such as one ALU = one lexical entry. NooJ provides linguists with the formal tools and methodology necessary for this formalisation. The dictionaries **DEM** (*Dictionnaire Electronique des Mots*)[9] and **LVF** (*Les Verbes Français*)[10] from Dubois & Dubois-Charlier could become the basis of an ideal electronic dictionary for NooJ.[11]

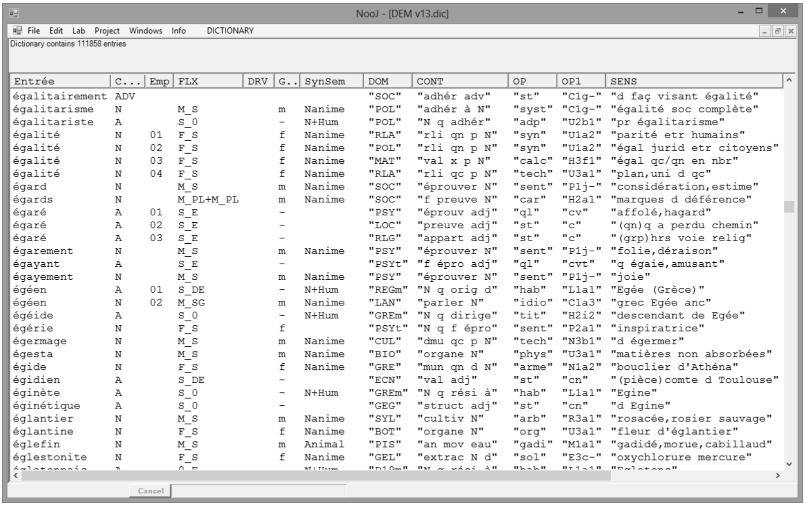
The DEM dictionary

The DEM dictionary (*Dictionnaire Électronique des Mots*) contains 145,135 entries in all morpho-syntactic categories.

[9] Cf. (Dubois 2010).

[10] Cf. (Dubois 1997).

[11] Cf. (Sabatier 2013).



Dictionary contains 111858 entries

Entrée	C...	Emp	FLX	DRV	G..	SynSem	DOM	CONT	OP	OP1	SENS
égalitairement	ADV						"SOC"	"adhér adv"	"st"	"C1g-"	"d faç visant égalité"
égalitarisme	N		M_S		m	Nanime	"POL"	"adhér à N"	"syst"	"C1g-"	"égalité soc complète"
égalitariste	A		S_0		-	N+Hum	"POL"	"N q adhér"	"adp"	"U2b1"	"pr égalitarisme"
égalité	N	01	F_S		f	Nanime	"RLA"	"rli qn p N"	"syn"	"U1a2"	"parité etr humains"
égalité	N	02	F_S		f	Nanime	"POL"	"rli qn p N"	"syn"	"U1a2"	"égal jurid etr citoyens"
égalité	N	03	F_S		f	Nanime	"MAT"	"val x p N"	"calc"	"H3f1"	"égal qc/qn en nbr"
égalité	N	04	F_S		f	Nanime	"RLA"	"rli qc p N"	"tech"	"U3a1"	"plan,uni d qc"
égard	N		M_S		m	Nanime	"SOC"	"éprouver N"	"sent"	"P1j-"	"considération,estime"
égards	N		M_PL+M_PL		m	Nanime	"SOC"	"f preuve N"	"car"	"H2a1"	"marques d déférence"
égaré	A	01	S_E		-		"PSY"	"éprouv adj"	"ql"	"cv"	"affolé,hagard"
égaré	A	02	S_E		-		"LOC"	"preuve adj"	"st"	"c"	"(qn)q a perdu chemin"
égaré	A	03	S_E		-		"RLG"	"appart adj"	"st"	"c"	"(grp)hrs voie relig"
égarement	N		M_S		m	Nanime	"PSY"	"éprouver N"	"sent"	"P1j-"	"folie,déraison"
égayant	A		S_E		-		"PSYt"	"f épro adj"	"ql"	"cvt"	"q égaie,amusant"
égayement	N		M_S		m	Nanime	"PSY"	"éprouver N"	"sent"	"P1j-"	"joie"
égéen	A	01	S_DE		-	N+Hum	"REGm"	"N q orig d"	"hab"	"L1a1"	"Egée (Grèce)"
égéen	N	02	M_SG		m	Nanime	"LAN"	"parler N"	"idio"	"C1a3"	"grec Egée anc"
égéide	A		S_0		-	N+Hum	"GREm"	"N q dirige"	"tit"	"H2i2"	"descendant de Egée"
égérie	N		F_S		f		"PSYt"	"N q f épro"	"sent"	"P2a1"	"inspiratrice"
égermage	N		M_S		m	Nanime	"CUL"	"dmu qc p N"	"tech"	"N3b1"	"d égermer"
égesta	N		M_S		m	Nanime	"BIO"	"organe N"	"phys"	"U3a1"	"matières non absorbées"
égide	N		F_S		f	Nanime	"GRE"	"mun qn d N"	"arme"	"N1a2"	"bouclier d'Athéna"
égidien	A		S_DE		-		"ECN"	"val adj"	"st"	"cn"	"(pièce)comte d Toulouse"
éginète	A		S_0		-	N+Hum	"GREm"	"N q rési à"	"hab"	"L1a1"	"Egine"
éginétique	A		S_0		-		"GEG"	"struct adj"	"st"	"cn"	"d Egine"
églantier	N		M_S		m	Nanime	"SYL"	"cultiv N"	"arb"	"R3a1"	"rosacée,rosier sauvage"
églantine	N		F_S		f	Nanime	"BOT"	"organe N"	"org"	"U3a1"	"fleur d'églantier"
églefin	N		M_S		m	Animal	"PIS"	"an mov eau"	"gadi"	"M1a1"	"gadidé,morue,cabillaud"
églestonite	N		F_S		f	Nanime	"GEL"	"extrac N d"	"sol"	"E3c-"	"oxychlorure mercure"

Figure 1 – The DEM dictionary

Each lexical entry is associated with a dozen properties, including: [12]

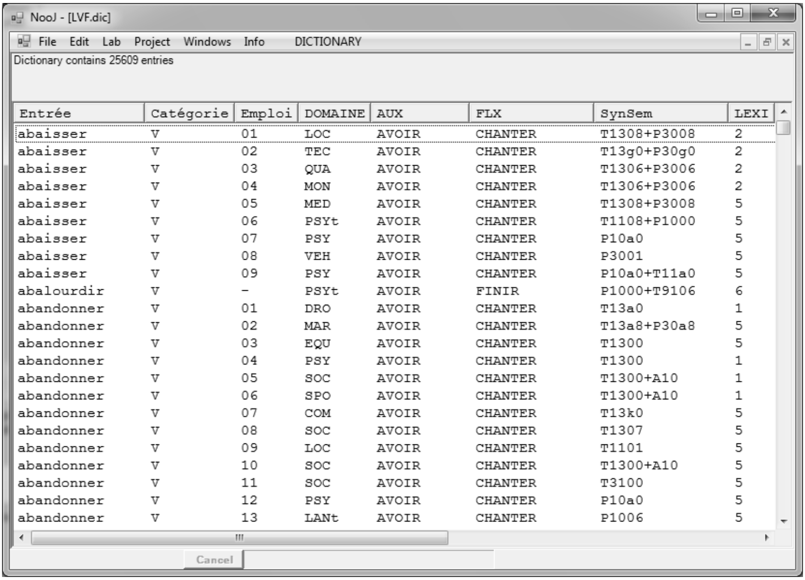
CAT (syntactic category) for instance Adverb, Verb, etc.
SENSE: each meaning is represented as a different lexical entry
DOMAIN: semantic domain of the term.
OPER: Semantic prototypical scheme of the term.
SCLASS: semantic class of the term.

Note that, as opposed to the DELA or the Lexicon-Grammar dictionaries, CAT may have more than one value, for instance when one ALU can be used both as a Noun and as an Adjective, eg *artiste*. This possibility is crucial if we want to satisfy the constraint '1 ALU = 1 lexical entry'. The SENSE, DOMAIN and SCLASS properties ensure that we always have '1 ALU= 1 lexical entry'.

The LVF Dictionary

The LVF dictionary (*Les Verbes Français*) contains 25,609 verbal entries.

[12] I translate each property code into English for better clarity.



Dictionary contains 25609 entries

Entrée	Catégorie	Emploi	DOMAINE	AUX	FLX	SynSem	LEXI
abaisser	V	01	LOC	AVOIR	CHANTER	T1308+P3008	2
abaisser	V	02	TEC	AVOIR	CHANTER	T13g0+P30g0	2
abaisser	V	03	QUA	AVOIR	CHANTER	T1306+P3006	2
abaisser	V	04	MON	AVOIR	CHANTER	T1306+P3006	2
abaisser	V	05	MED	AVOIR	CHANTER	T1308+P3008	5
abaisser	V	06	PSYt	AVOIR	CHANTER	T1108+P1000	5
abaisser	V	07	PSY	AVOIR	CHANTER	P10a0	5
abaisser	V	08	VEH	AVOIR	CHANTER	P3001	5
abaisser	V	09	PSY	AVOIR	CHANTER	P10a0+T11a0	5
abalourdir	V	-	PSYt	AVOIR	FINIR	P1000+T9106	6
abandonner	V	01	DRO	AVOIR	CHANTER	T13a0	1
abandonner	V	02	MAR	AVOIR	CHANTER	T13a8+P30a8	5
abandonner	V	03	EQU	AVOIR	CHANTER	T1300	5
abandonner	V	04	PSY	AVOIR	CHANTER	T1300	1
abandonner	V	05	SOC	AVOIR	CHANTER	T1300+A10	1
abandonner	V	06	SPO	AVOIR	CHANTER	T1300+A10	1
abandonner	V	07	COM	AVOIR	CHANTER	T13k0	5
abandonner	V	08	SOC	AVOIR	CHANTER	T1307	5
abandonner	V	09	LOC	AVOIR	CHANTER	T1101	5
abandonner	V	10	SOC	AVOIR	CHANTER	T1300+A10	5
abandonner	V	11	SOC	AVOIR	CHANTER	T3100	5
abandonner	V	12	PSY	AVOIR	CHANTER	P10a0	5
abandonner	V	13	LANt	AVOIR	CHANTER	P1006	5

Figure 2 – The LVF dictionary

Each entry of the LVF dictionary is associated with 10 properties, including:

SENSE: each meaning is represented as a different lexical entry
DOMAIN: semantic domain for the verb.
OPER: semantic prototypical scheme of the verb.
SCLASS: semantic class of the verb.
CONJUGATION: the conjugation paradigm for the verb.
STRUCTURE: one or more syntactic structures for the verb.
DERIVATION: potential derivational paradigms for the verb.

The fact that each verbal entry is associated with inflectional, derivational, syntactic and semantic information imposes the use of various processing tools, which is compatible with the NooJ approach, but not with ‘mono-formalism’ approaches. Here is for instance a lexical entry for the NooJ dictionary based on the DEM and LVF dictionaries:

abaisser,
V+SENSE=01+DOMAIN=LOC+AUX=AVOIR+FLX=CHANTER
+SCLASS=T3c+T1308+P3008+LEVEL=2+DRV=BASE1
+DRV=ABLE+DRV=MENT+DRV=EUR+OPER='r/d bas qc'

abaisser is a verb (V); the lexical entry corresponds to sense #1 (SENSE=01); its semantic domain is 'Locative' (DOMAIN=LOC); its semantic class is 'T3c'; its semantic analysis is 'r/d bas qc' (make something low). The verb is conjugated according to the paradigm 'CHANTER' and is conjugated with auxiliary verb *avoir* (AUX=AVOIR). It accepts the two following syntactic structures: T3108 (Human Subject + Verb + Non-animated Object + Instrumental Complement) and P3008 (Non-animated Subject + Pronominal Verb + Instrumental Complement). The verb accepts three derivations: the adjective *abaissable* (DRV=ABLE), the noun *abaissement* (DRV=MENT) and the noun *abaisseur* (DRV=EUR). The verb is associated with the base noun *abaisse*. It belongs to the basic French vocabulary (LEXI=2).

The reader will appreciate the huge qualitative difference with the DELA dictionaries. As an example, I now present how I used the **STRUCTURE** syntactic property in the LVF[13] to construct a new type of search engine capable of finding specific meanings for a given verb.

The STRUCTURE property

Each of the lexical entries of the LVF dictionary is associated with one or more syntactic structures, among 318 different ones[14], that I have described using four generic grammars: A (intransitive structure), N (indirect transitive structure), P (pronominal structure), and T (direct transitive structure). Each of the A, N, P and T structures has been implemented by a corresponding NooJ grammar.

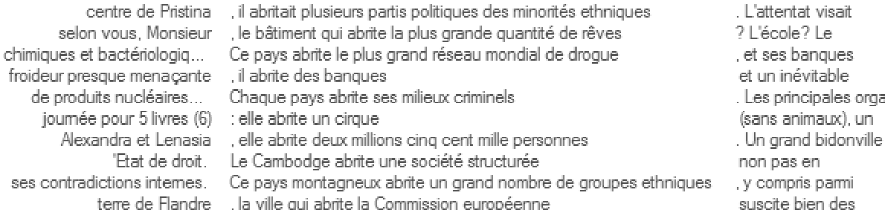
Remember that when a verb has more than one meaning, each of its meanings is listed as a separate entry, and is associated with its syntactic structures. For instance, consider the five senses for the verb *abriter* as described in the LVF dictionary:

[13] (Silberztein 2010) presents a more detailed description of the process of adapting the LVF dictionary so that NooJ parsers can process its information.

[14] (François, Le Pesant, Leeman 2007) present the semantic classification of the LVF dictionary in detail.

abriter #1 (T11b8, P10b8): *Luc abrite Léa de la pluie avec son parapluie*
‘Luc protects Lea from the rain with his umbrella’
abriter #2 (T1101, P1001): *Luc abrite des réfugiés chez lui*
‘Luc gives shelter to refugees’
abriter #3 (T3100): *Cet immeuble abrite les services du Ministère de l’éducation*
‘This building hosts the Minister of Education’
abriter #4 (P10b1): *Luc s’abrite des ennuis derrière Léa*
‘Luc hides from trouble behind Lea’
abriter #5 (T13b8): *Luc abrite le port des vagues avec des digues*
‘Luc shields the harbour from the waves with a seawall’

Thanks to the LVF dictionary as well as the four grammars A, N, P and T, NooJ can extract all the sentences from a corpus of texts that contain a specific structure for a given verb, and hence identify its specific meaning. For instance, by applying grammar T to a corpus constituted by over 7,000 articles of the newspaper *Le Monde diplomatique*, NooJ produces the following concordance:

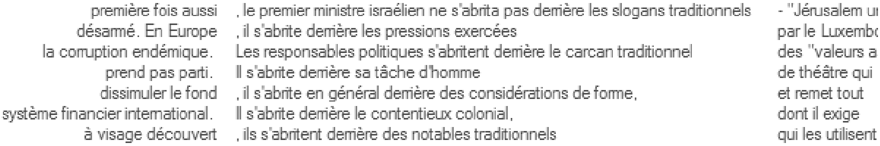


```
       centre de Pristina   , il abritait plusieurs partis politiques des minorités ethniques      . L'attentat visait
     selon vous, Monsieur   , le bâtiment qui abrite la plus grande quantité de rêves              ? L'école? Le
chimiques et bactériologiq...  Ce pays abrite le plus grand réseau mondial de drogue               , et ses banques
 froideur presque menaçante   , il abrite des banques                                               et un inévitable
   de produits nucléaires...  Chaque pays abrite ses milieux criminels                             . Les principales orga
    journée pour 5 livres (6)  : elle abrite un cirque                                              (sans animaux), un
       Alexandra et Lenasia   , elle abrite deux millions cinq cent mille personnes                . Un grand bidonville
             'Etat de droit.  Le Cambodge abrite une société structurée                            non pas en
  ses contradictions internes.  Ce pays montagneux abrite un grand nombre de groupes ethniques     , y compris parmi
          terre de Flandre   . la ville qui abrite la Commission européenne                         suscite bien des
```

Figure 3 – Occurrences of the verb *abriter* used in a direct transitive structure

We find occurrences of sense #3 of the verb *abriter*, whereas senses #1, #2 and #5 are excluded since they require prepositional complements that are not present in the sentences[15]. In the same way, if we apply grammar P, NooJ produces the following concordance:

[15] The N, P and T grammars contain the same reference to prepositional noun phrases. In the original LVF dictionary, the ‘b’ character in structure codes T11b8 and P10b8 represents the *de* preposition.



```
        première fois aussi   , le premier ministre israélien ne s'abrita pas derrière les slogans traditionnels   - "Jérusalem ui
         désarmé. En Europe   , il s'abrite derrière les pressions exercées                                        par le Luxembo
   la corruption endémique.   Les responsables politiques s'abritent derrière le carcan traditionnel             des "valeurs a
           prend pas parti.   Il s'abrite derrière sa tâche d'homme                                               de théâtre qui
         dissimuler le fond   , il s'abrite en général derrière des considérations de forme,                       et remet tout
système financier international.   Il s'abrite derrière le contentieux colonial,                                   dont il exige
         à visage découvert   , ils s'abritent derrière des notables traditionnels                                 qui les utilisent
```

Figure 4 – Occurrences of the verb *abriter* in a pronominal structure

Here, we find occurrences of sense #2, whereas senses #1 and senses 4 are excluded because there is no complement introduced by the preposition *de*. Therefore, NooJ can extract sentences from a corpus of texts that contain a specific meaning for a verb: no other search engine or linguistic corpus processor can perform this type of operation. This application constitutes qualitative progress in corpus linguistics.

Distribution Selection

The STRUCTURE code includes the distributional class of each verb argument. For instance, the two first digits correspond to the subject and to the object complement of the verb: ‘1’ = Human; ‘3’ = Thing. The last character corresponds to the specific adverbial complement (the ‘*circonstant’*): ‘0’ = no complement, ‘1’ = locative complement, ‘8’ = instrumental complement. The distributional selection described in the LVF dictionary is useful in order to distinguish between different meanings of a given verb, and they can be used to link verbs to the nouns that are described in the DEM dictionary. For instance, in order to distinguish between sense #2 and sense #3 of the verb *abriter*, we can use the fact that sense #2 requires a Human subject, whereas sense #3 requires a locative subject. However, adding distribution selection constraints in the four generic grammars would stop NooJ from finding a number of occurrences for each meaning, for several reasons:

— Metonymies (which are frequent) would be systematically rejected. For instance, the French word ambassade (embassy) is described in the DEM dictionary as a non-human noun. In consequence, sentences such as ‘The embassy invited the Prime Minister’ would not be recognised, since the word ‘embassy’ here plays the role of a Human noun.

— More generally, a large number of noun phrases that are ‘lexically non-human’ in fact play the role of humans in certain contexts. For instance, the sentence Luc achète le silence du fonctionnaire (Luc buys the government employee’s silence) is associated with the structure T1106 (object complement = Human); however the noun silence cannot be described as Human in the DEM dictionary.

Inversely, a number of human noun phrases can acquire the function of a thing, depending on their context. For instance, sense #1 of the verb *agacer* in *Ces nouvelles agacent Luc* (this news bothers Luc) is associated with the structure code T3100 (subject = thing). But it is totally possible to have a human noun as a subject in this sentence[16], eg in *Léa agace Luc* (Lea bothers Luc). In other words, it is possible to use the distribution selection information provided in the LVF dictionary, linked with the corresponding distributional class described in the DEM dictionary: the resulting concordance would be more precise, but at the expense of a much lower recall. I believe that this is still the right approach: sentences that have been 'rightfully' rejected by the NooJ grammars should be the objects of some linguistic computation capable of solving metonymies and metaphors.

Conclusion

Our first experiments show how rich the DEM and the LVF dictionaries are. Using the information stored in these two dictionaries can greatly enhance the precision of NLP software applications, by shifting from 'word-based' to 'sense-aware' search engines. Transforming and enriching these two dictionaries in order to construct the electronic dictionary that will formalise the French vocabulary would probably take several years for a small team of linguists:

— We will need to fully merge these two dictionaries. Most verbs described in the DEM are also described in the LVF; some verbs in DEM are not described in the LVF dictionary. A number of nouns and adjectives are represented both as independent entries in the DEM and as derived from a verbal entry in the LVF.

— The entries of the DEM dictionary have no inflectional or derivational description. We will have to add both descriptions.

— Several codes in the DEM and LVF dictionaries encode combinations of properties that must be properly untangled and formalised. For instance, property CAT in the DEM represents three types of information: the morpho-syntactic category (eg

[16] (Gross 1975) uses the 'unrestricted' category to describe complements that can accept any noun, such as the subject of sentence: *(Luc | This country | The table | The dog | This event | The rain | The fact that she came) bothers Lea*.

Noun), the gender (eg Feminine) and the semantic class (eg Human)[17].

— We will need to add to these dictionaries multiword units, frozen expressions as well as Support Verb/Predicative Noun combinations[18].

References

Courtois, Blandine. 1990. Un système de dictionnaires électroniques pour les mots simples du français. In *Dictionnaires électroniques du français*, Courtois, Silberztein Eds. Langue française n° 87, pp. 5-10. Larousse : Paris.

Courtois, Blandine and Max Silberztein Éds. 1990. Les dictionnaires électroniques. Langue française n° 87. Paris : Larousse.

Dubois, Jean and Françoise Dubois-Charlier. 1997. Les verbes français.

—. 2010. La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. In *Langages 179-180.* Armand Collin : Paris, pp. 31-56.

Fairon, Cédric éd. 1999. Analyse lexicale et syntaxique : le système INTEX. Lingvisticae Investigationes, tome XXII: 1998-1999. John Benjamins Publishing Company: Amsterdam.

François, Jacques, Denis Le Pesant, and Danielle Leeman Eds. 2007. Le classement syntactico-sémantique des verbes français. Langue française n° 153. Armand Colin : Paris.

Gross, Maurice. 1975. *Méthodes en syntaxe*. Hermann : Paris.

—. 1977. Grammaire transformationnelle du français, 2 : Syntaxe du nom. Larousse : Paris.

Leclère, Christian. 1990. Organisation du lexique-grammaire des verbes français. In *Dictionnaires électroniques du français*, Courtois, Silberztein Eds. Langue française no 87, pp. 104-111. Larousse : Paris.

[17] I have refined the inflectional description of the LVF dictionary because certain conjugation paradigms of LVF had to be separated into several different NooJ paradigms; conversely, several different conjugation paradigms could be unified thanks to the use of NooJ ‘intelligent’ morphological operators, cf. (Silberztein 2010b). Paradigms in LVF did not take defective or impersonal conjugations into account: I had to add the corresponding paradigms in NooJ.

[18] A large number of frozen expressions and Support verb/Predicative nouns combinations have been listed and described in Lexicon-Grammars, but, here too, we will need to add missing properties in order to import them into the NooJ platform.

—. 1998. Travaux récents en lexique-grammaire. Travaux de linguistique n°37. Rijksuniversiteit van Gent Ed. p 191

Sabatier, Paul and Denis Le Pesant. 2013. Les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier et leur exploitation en TAL. In *Ressources Lexicales*, *Linguisticae Investigationes Supplementa* 30. John Benjamins Publishing Company : Amsterdam.

Silberztein, Max. 1990. Le dictionnaire électronique des mots composés. In *Dictionnaires électroniques du français*, Courtois, Silberztein Eds. Langue française n° 87, pp. 11-22. Larousse : Paris.

—. 1993. Dictionnaires électroniques et analyse automatique de textes : le système INTEX. Masson : Paris.

—. 2004. Une description formalisée des déterminants français. In *Hommage à la mémoire de Maurice Gross*. Linguisticae Investigationes, E. Laporte, C. Leclère, M. Piot, M. Silberztein Eds. pp. 589-600.

—. 2005. NooJ Dictionaries. In *Proceedings of the 2nd Language and Technology Conference*. Poznan.

—. 2010. La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. In *Théorie, empirie, exploitation : l'exemple des travaux de Jean Dubois sur les verbes français*. Langages n° 179-180, Danielle Leeman, Paul Sabatier Eds.

Trouilleux, François. 2012. A New French Dictionary for NooJ : le DM. In *Selected papers from the 2011 International NooJ Conference*. Cambridge Scholar Publishing : Newcastle.

THE FORMALISATION OF MOVEMENT VERBS FOR AUTOMATIC TRANSLATION USING NOOJ PLATFORM

HAJER CHEIKHROUHOU

Abstract

This paper is concerned with French verbs and in particular, the movement verbs (entry/exit). In this paper, we will first propose a semantic and syntactic description of the movement verbs while relying on Dubois's dictionary. Second, we will show the linguistic characteristics of the Arabic verbs. Finally, we will try to use the platform NooJ to achieve an automatic translation of movement verbs.

Introduction

The mass of documents has become increasingly difficult to operate and manage. In fact, the user encounters many difficulties in finding the relevant information, especially if it is not written in the preferred language. As a result, a new need, regarding translation of this information to the desired language, has emerged.

Thus, the requirement for more reliable automatic translation systems is increasing. For this reason, we are interested, in this study, in devising a machine translation system for the French-Arabic language pair. We chose mainly to process and analyse the verb since it is a fundamental element in the structure of sentences in all natural languages. In this context, we will essentially study the movement verbs (entry/exit), which constitute the E class database of Jean and Françoise Dubois - Charlier (LVF).

In this paper, we will first propose a semantic and syntactic description of the movement verbs relying on Dubois's dictionary. Secondly, we will show the linguistic characteristics of the Arabic verbs. Finally, we will try to use the NooJ platform to achieve an automatic translation of movement verbs.

The Linguistic Description of the E Class

The French Verbs of Jean Dubois and Françoise Dubois-Charlier (LVF) is a thesaurus of syntactic-semantic classes. The LVF is composed of 25,610 entries for 12,310 different verbs. There are fourteen classes, including Class E which contains 2,444 entries representing the class of movement verbs. To accomplish this task, we will examine the semantic-syntactic classes, operators, syntactic constructions and, finally, the domain.

The semantic-syntactic classes

Class E contains four semantic-syntactic classes E1, E2, E3 and E4.

E1	-«sortir/venir de qp, aller/entrer qp», sujet humain ou animal propre (« leave/come from somewhere, go/enter into somewhere» human subject or animal subject) -«faire sortir, aller, entrer qnqp» (« let out, go, enter s.o somewhere»)	7 subclasses
E2	- figuré de E1 (figurative E1)	5 subclasses
E3	- «sortir/venir de qp, aller/entrer qp», sujet non-animé («leave/come from somewhere, go/enter somewhere» inanimate subject) -«faire sortir/entrer qc qp» («let out/in s.o somewhere»)	6 subclasses
E4	- figuré de E3 (figurative E3)	6 subclasses

Table 1 – Semantic and syntactic classes of Class E

The Syntactic subclasses

The four semantic-syntactic classes are divided into tewenty-four syntactic subclasses.

<table>
<tr><td colspan="3"><b>Class E1, 698 entries, «sortir/venir de qp ; aller/entrer qp», sujet humain ; «faire sortir, aller, entrer qn qp».</b>
(« leave/come from somewhere, go/enter into somewhere « human subject , « let out/in, go, enter s.o somewhere»)</td></tr>
<tr><td><b>Syntactic sub-classes</b></td><td><b>Entries</b></td><td><b>Example</b></td></tr>
<tr><td>E1a-b-c-d-e-f-g</td><td>698entries</td><td>Aller15(Go), Attirer03(Attract)</td></tr>
<tr><td colspan="3"><b>Class E2, 440 entries, « figuré de E1 » humain figuré</b>
(« figurative E1 » human figurative)</td></tr>
<tr><td>E2a-b-c-d-e</td><td>440entries</td><td>Avancer04(Advance),Balancer06(Sway)</td></tr>
<tr><td colspan="3"><b>Class E3, 984 entries, «(faire) sortir/venir de qp ; (faire) aller/entrer qp», sujet non-animé propre</b>
(«leave/come from somewhere, go/enter somewhere» inanimate subject)</td></tr>
<tr><td>E3a-b-c-d-e-f</td><td>984 entries</td><td>Vibrer03 (Vibrate), Courir03 (Run)</td></tr>
<tr><td colspan="3"><b>Class E4, (322 entries), «figuré de E3», sujet non-animé figuré</b>
(«figurative E3 » inanimate figurative)</td></tr>
<tr><td>E4a-b-c-d-e-f</td><td>322 entries</td><td>Sortir10 (Leave),Venir06 (Come)</td></tr>
</table>

Table 2 – The syntactic subclasses of E1

The operators

Each verb entry is defined with a syntactico-semantic diagram, encoded with a sequence of alphabetical characters called operator. Class E operators are:

ex = sortir de (leave)	**f.ex = faire sortir de (leave)**	**f.ire = faire aller qp (go to)**	**ire = aller qp go somewhere**

The syntactic constructions

The coding of syntactic constructions includes a combination of letters and numbers such as the coding [N3b] which means that the verb is transitive with an indirect complement introduced by the preposition «de». **Example:** *dériver* (06) (derive), *émigrer* (03) (migrate).

We notice that there are verbs that take two syntactic constructions like the word *débarquer* (disembark). The latter belongs to the subclass syntax E1b and can have two syntactic constructions A13 and T1130.

- A13 means that the verb is intransitive with a human subject and a locative complement.

- T1130 means that the verb is transitive with a human subject, a human direct object and a locative complement.

Operators that define this sub-syntactic classification are related to the operator f.ex which means *faire sortir* (leave).

The Domain

Dubois, in LVF, devotes a section to the field (DOM) which states:

- the pragmatic, technical and scientific verbal domains
- the levels of language and regionalism

Concerning the pragmatic domains to which the movement verbs belong, we essentially have:

- *Locatif et lieux* (locatives, places) coded by **LOC**. These verbs represent the majority of the verbs of movement. The verbs that belong to this domain can be at the familiar, popular, literary and old language level.

Examples:

— *Abattre*(09) (shoot) belongs to the domain LOC ie locative.
— *Balader*(01) (backpack) is a familiar register encoded by LOCf.
— *Carapater*(s) (carapate) is a popular register encoded by LOCp, the meaning of the verb is *s'enfuir (*to escape).
— *Cortéger* (accompany) is a literary register coded LOCt whose meaning is *accompagner, escorter (*to accompany, escort).
— *Ensauver*(s) is encoded by an old register Locv whose meaning is *s'enfuir, se débiner (*to flee, to run away).

— ***Elevage*** (Elevation) is coded by **ELV**.

Examples:

— *Chasser*(05) (hunt) = *pousser devant soi (*push ahead)
— *Traire*(02) (milk) = *tirer le lait de* (shoot milk)

●

— ***Bâtiment*** (building) encoded by **BAT**.

Examples:

— *Cloisonner01* (partition) = *comprtimenter* (compartmentalise)
— *Décoffrer* (stripping framework) = *ôter de son coffrage* (remove from its casing)

The linguistic characteristics of the Arabic verb

For grammarians of Arabic, the verb is an essential element in the construction of the sentence. Associated with the subject, it forms the core of the sentence. Around this core, other items are ordered. For this reason, it is classified as a basic component. In the most important dictionary of Arabic, *lisān al -ʻarab de Ibn Manẓū r*, we found the following definition of the verb (*fi'l*):

> 'al- fiʻlu kināyatun ʻan kulli ʻamalin mutaʻaddin ʼaw ġayri mutaʻaddin (…) maṣdar min faʻala yafʻalu - faʻlan wa fiʻlan '[1]. (The verb is the name of any do transitive or intransitive (...) it is a name of action fa'ala - yaf'alu*)*

In fact, in grammar, the oldest definition of the verb dates back to Sibawayhi. It distinguishes the concept of the verb and the name of the particle. He says in the chapter (bābu ʻilmi mā al- kalimu mina -l- ʻarabiyya):

> ʻ … wa ʼammā al-fiʻlu faʼamṯilatun ʼuḫiḏ at min lafẓi ʼaḥdāṯi al-ʼasmāʼi wa buniyat limā maḍā wa limā yakūnu wa lam yaqaʻ wa mā huwa kāʼinun lam yanqaṭiʼ (As for verbs, they are structures derived from the noun and built on what happened, what will happen or be and what did not happen, and what is and what is not interrupted).

This definition emphasises the two dimensions that comprise the verb, namely action and time.This implies that any verb is derived from the name of the action *''ismu al-ḥadaṯi'* which is nothing other than the (*maṣdar*). As for the concept of time, Sibawayhi did not merely detail it. He clarified that the verb (*fi'l*) is either past or present or future.

Like Hebrew and Syriac, Arabic is a Semitic language. In this family of languages, we give the term ʻconsonantal root*'* to the consonant clusters that occur in a fixed order. The number of consonants, called the radical consonants, is:

- **Three**: in this case we speak of a triconsonantal root.
  eg: دخل (to enter)
- **Four**: in this case we speak of a quadriconsonantal root.
  eg: دحرج (to budge)

[1] Ibn Manẓūr : Lisān ʼal -ʻarab : terme (faʻala) (1956 : XI, p. 568).