LE DM, A FRENCH DICTIONARY FOR NOOJ

FRANÇOIS TROUILLEUX

Abstract

This paper presents the DM, a new dictionary for French. Freely available resources are selectively used to obtain lexical lemmas, from which morphological grammars generate about 538 000 baseforms. Evaluation of the DM on corpus shows that it stands the comparison with the previous NooJ delaf dictionary.

Introduction

For historical reasons, large coverage French dictionaries are available to NooJ users only in the compiled *.nod* format. This poses several problems for grammar development, e.g. constraints won't work, adding new information to a lexical entry requires redefining the whole set of information on that entry, the dictionaries cannot be used for generation…

We then decided to produce a new dictionary, called the DM, designed in the NooJ format (Silberstein 2003, 2005), on which the NooJ community will have control. This paper first presents a quantitative analysis of the freely available resources we considered. The next two sections describe the way we constructed the DM for lexical and function words, leading to a global view of the DM extension. Finally, results of morphological analysis and parsing with the DM are compared to results obtained with the *delaf.nod* dictionary.

Available resources

To build a freely available dictionary, we had to rely on resources for which free reuse is licensed. This excludes some resources, in particular the Morfetik dictionary (Mathieu-Colas 2009), which is interesting in that it makes use of several good quality resources. Our initial plan was to rely on three free resources: the DELA (Courtois 1990), Morphalou (Romary *et al.* 2004) and the Le*fff* (Sagot 2010)[1]. In order to evaluate the potential contribution of each of these resources, we compared the lemmas they contain for the adjective, adverb, common noun, verb, interjection and prefix categories. Results are given in Table 1[2]. The *union* column gives the number of lemmas in the union of the three dictionaries. The central columns give the percentage of the union which is common to the three dictionaries (*int.*), common to only two (*DL, LM, MD*) and specific to each (*D, L, M*).

	int.	**DL**	**LM**	**MD**	**D**	**L**	**M**	**union**
adjectives	24.7	5.1	2.5	15.4	23.1	11.6	17.6	35124
adverbs	58.9	22.9	0.9	0.8	8.1	2.9	5.4	3704
nouns	35.5	6.1	0.3	9.4	30.8	1	16.9	82118
verbs	45.7	3.5	0.1	12.2	29.1	1.9	7.5	13271
interj.	3.4	2.3	0.3	30.1	23	10.5	30.4	352
prefixes	0	9.6	0	0	84.1	6.3	0	921

Table 1. DELA-Lefff-Morphalou comparison.

	int.	**dela**	**morph.**	**union**
adjectives	45.3	32	22.7	31064
adverbs	61.6	32	6.5	3596
nouns	45.4	37.3	17.4	81289
verbs	59	33.2	7.8	13020
interjections	37.5	28.3	34.3	315

Table 2. DELA-Morphalou comparison.

The prefix and interjection categories are peculiar. There are no prefixes as autonomous entries in Morphalou. For this category, the DELA is much richer than the Le*fff*, the intersection of the two being rather small, with only 88 lemmas. Interjections are dealt with disparately in the three dictionaries. The table line actually counts the following categories: *intj* for the DELA, *interjection* and *onomatopoeia* for Morphalou and *pres* for the Le*fff*. As can be seen, the intersection is not empty, but the categories do not correspond very well. 23 onomatopoeia of Morphalou are

[1] First work on the Le*fff* dates back to 2004; we use the latest version to date, extensional version 3.0.

[2] We only look at uncapitalized *simple* words, *i.e.* without any whitespace, hyphen nor apostrophe. Figures are obtained after correction of a few errors and normalization of some lemmas.

interjections in the DELA; the category *pres* of the Le*fff* includes both interjections and presentatives as *voici*, *voilà* ("here is"), which insofar as they may introduce a complement and combine with clitic pronouns should maybe be categorized differently.

Regarding the four other, more important, categories, one may note that the specific contribution of the Le*fff* is small, except for adjectives. More than 96% of the adverbs, nouns and verbs of the Le*fff* are actually present in the DELA. For adjectives, the number is only 67%, but the difference comes from the fact that the Le*fff* quite systematically codes past participles as adjectives: the "adjectives" *exhumé*, *blasphémé*, *démarré* ("started"), for instance, are specific to the Le*fff*.

The specific contribution of Morphalou is relatively small for adverbs and verbs. This is due to recent work on these two categories in the Le*fff* (cf. Sagot and Fort 2007; Tolone and Sagot 2009). On the other hand, Morphalou contains an important number of nouns and adjectives which are neither in the DELA nor the Le*fff*.

Table 2 gives a direct comparison between the DELA and Morphalou. It shows that the intersection of the two dictionaries is surprisingly small and that the DELA is clearly the bigger of the two dictionaries.

The DM Lexical Words

Lemma Selection

In view of the observations we made, we had several options for a new dictionary. A first idea could be to make the union of the three dictionaries and obtain what would probably be the biggest freely available French dictionary for NLP. The drawback of this approach would be that the dictionary will include all the errors to be found in the dictionaries. Rather than taking the union of the dictionaries, we then decided to go for the intersection. The idea is to favor precision: the presence of a word in several dictionaries is a guarantee that it does exist in French. Each dictionary is in a way validated against the others. With the choice of the intersection, this project differs from the Morfetik project (Mathieu-Colas 2009), which builds the *union* of the resources it uses; the choice of the union in this project imposes manual validation of the entries, our choice of the intersection allows automatic validation.

Having chosen to favor precision, the question remains which intersection to take. A first idea could be to take the intersection of the three dictionaries, but, as we have seen, the Le*fff* is for a very large part included in the DELA. The DELA would thus have a sort of double weight in the intersection of the three. We then decided to use the intersection of the DELA and Morphalou only to produce the DM.

In addition to the fact that the intersection will produce a higher quality result, one may also note other motivations for this choice: a smaller dictionary will offer an opportunity to test the influence of the dictionary size on parsing results, it will be easier to check, and it will always be possible to add new entries in the future.

In the end, the DM dictionary has been built according to the following procedure: (1) select the lemmas in the intersection of the DELA and Morphalou, plus all the prefixes of the DELA; (2) couple these lemmas with NooJ morphological grammars to generate the inflected forms.

As mentioned above, the set of lemmas is limited to words without any whitespace, hyphen nor apostrophe.

Generation of inflected forms

The two dictionaries do not always give the same set of forms for the same lemma. For instance, for *dandy,* the DELA gives one plural form: *dandys,* while Morphalou gives two: *dandys*, *dandies*. For nouns which are also adjectives, Morphalou often gives more forms than the DELA, usually four, e.g. for people names (e.g. for *français*, the DELA gives one form, while Morphalou gives four) and for some adjectives which as masculine singular nouns denote the quality specified by the adjective (e.g. *arbitraire,* meaning « arbitraryness »). We used the forms of the DELA as a reference. The forms generated by the DM grammars have been systematically compared to those of the DELA, so that the reliability of the grammars is guaranteed.

This, however, does not forbid a few differences. For instance, the grammars are designed to systematically generate French style plural for foreign words (e.g. *maffiosos* and *maffiosi*, *recordmans* and *recordmen*, *corpus* and *corpora*, etc.). For verbs, the grammars overgenerate some inflected forms for impersonal verbs (e.g. *je faudrai*) and feminine or plural past participles for intransitive verbs. That the dictionary should or should not generate possible but unattested forms is an open question.

Morphological Grammars

Inflected forms for nouns, adjectives and verbs are generated by NooJ grammars. There are 109 flexion paradigms for nouns and adjectives. Regular paradigms are named according to a regular coding scheme:

- M or F (“masculine” or “feminine”), for words marked in gender;

- the characteristic mark of the plural (0 if there is no such mark), except for nouns in *–al* and *–ail*, plural *–aux*, for which the singular characteristic mark is used;
- the characteristic mark of a second plural, if any; the regular plural is indicated first (e.g. M_MANS_MEN, a paradigm which generates *tennismans* and *tennismen*);
- the characteristic mark of the feminine singular (0 if unmarked), except for words which double the final consonant (code: DE).

As an illustration, Table 3 gives the 18 most frequently used paradigms for nouns and adjectives, with the number of times they are used. One can see that the two most used paradigms are M_S and F_S, *i.e.* two forms marked in gender, with plural by adding an *s*[3]. Next come two paradigms with four forms: S_0 (add an *s* for plural, nothing for feminine, e.g. *troisième*) and S_E (add an *s* for plural, an *e* for feminine, e.g. *petit*). This naming scheme for flexion paradigms proved quite useful during development, because it appeals less to memory than the notation by a member of the class (e.g. PETIT instead of S_E) which is suggested in the NooJ documentation.

F_S	14851	M_0	975	M_SG	439
M_S	14778	AUX_ALE	686	M_X	295
S_0	7659	EUX_EUSE	608	0_E	175
S_E	4133	ERS_ERE	504	0_0	147
EURS_EUSE	1098	EURS_RICE	502	M_AL	65
S_DE	998	FS_VE	461	F_0	50

Table 3. 18 most frequently used noun and adjective paradigms.

Specifying a naming scheme for verb paradigms is not so obvious, so we decided to stick to the use of representatives such as AIMER, CROIRE, etc., which is also that of conjugation manuals. The DM makes use of 95 paradigms for verbs.

The interesting point to note is that the description of conjugation is structured. For the first group verbs, the endings are grouped into three classes: future and conditional tense endings (e.g. *chantera*), unstressed (closed final syllable, e.g. *chante*) and stressed (open final syllable, e.g. *chantez*). Stressed endings are further divided into two subclasses

[3] Technically, one could collapse these two paradigms into one and mark gender at the entry level; this would not bring any gain in terms of typing, however.

depending on their initial vowel (to account for variations of the stem such as *commençons* vs *commencions*). For the second and third group, the structuration is inspired by (Le Goffic 1997), which shows that, give or take a few exceptions, all the verbs may be described using at most six different stems (plus the infinitive). The grammar thus defines sets of endings which always share the same stem and these definitions are systematically used in the flexion paradigms themselves. For instance, in the VOIR paradigm reproduced below, one refers back successively to the endings of (1) the imperative and present indicative singular (stem *voi*), (2) the imperfect and two present subjunctive forms (stem *voy*), (3) the rest of the present subjunctive forms (stem *voi*), (4) the simple past and imperfect subjunctive forms (stem *v*), (5) the future and conditional forms (stem *verr*), and (6) the past participle forms (stem *vu*).

```
VOIR = (<B> :PRES-s-t ) | (<B2>y :IMP) | (<B> :SUBJ)
 | (<B3> :PS-IS-i) | (<B3>err :FC) | (<B3>u :PP) | <E>/INF ;
```

As complex as it seems at first sight, this description proved quite robust (errors are quickly spotted, since the endings are widely shared) and saved a lot of copy/paste text.

The DM Function Words

We did not use the DELA and Morphalou for function words as the work on the chunker for French presented in (Trouilleux 2009a, 2009b) already led us to define a dictionary of function words, using as a starting point data developed at Université Blaise-Pascal in the 90s. Consulting the DELA and Morphalou would be of little interest for two reasons. First, the sets of function words are closed and there should be few missing entries as the chunker has been evaluated on corpus. Second, the development of a parser often leads one to reconsider the traditional definition of some function words. Morphalou has been derived from the *Trésor de la langue française* (*TLF*), a traditional dictionary, designed with no concern for NLP; the DELA has been defined at a time when robust, large corpus parsers did not exist; the Le*fff*, on the contrary, is associated with a wide-coverage text parser (cf. Tolone and Sagot 2009). These differences sometimes may justify differences in the categorization of word forms.

As an example of this problem, Table 4 shows how a few function words are categorized in the DELA, the TLF[4], the Le*fff* and the DM (ignoring lexical categories). There are three cases.

		DELA	**TLF**	**Le*fff***	**DM**
1	*ne*	Adverb	adv.	clneg	ADV+neg+clit
2	*combien*	Adverb	adv.	pri	ADV+int
	comment	adverb conjs	adv.	pri	ADV+int
	où	pronoun	adv. pron.	pri prel	ADV+int PRO+rel
	pourquoi	adverb conjs	adv.	pri	ADV+int
	quand	adverb conjs	adv. conj.	pri csu	ADV+int CONJS
3	*avec*	Prep	prép.	prep adv	PREP+ell
	depuis	prep adv	prép. adv.	prep adv	PREP+ell
	pendant	prep adv	prép.	prep adv	PREP+ell
	pour	Prep	prép.	prep	PREP+ell

Table 4. Variations in the categorization of function words.

First, the negation adverb *ne* has a distribution which it shares with no other adverb. This justifies a specific category in the Le*fff*, additional features in the DM ("negation", "clitic").

Second, the DM contains five forms categorized as ADV+int (for "interrogative"). They have in common the capacity to introduce a direct or indirect interrogative clause. Among them, *où* ("where") may also introduce a relative clause and *quand* ("when") a non interrogative clause, but it is not the case for *comment* ("how") and *pourquoi* ("why"). One may thus note two faulty categorizations as *conjs* in the DELA, and an incomplete specification for *où*.

Third, the DM categorizes as PREP+ell (for "ellipsis") thirteen prepositions which allow the ellipsis of their complement. Our sample shows how disparately these forms are categorized. To us, they are fundamentally prepositions and the so called "adverb" readings are but object drop uses (cf. Trouilleux 2009b).

[4] Morphalou has only a generic *functionWord* category for prepositions, determiners, pronouns, etc.

Table 5 gives the list of categories used in the DM. The category codes are those of the DELA, except for three new categories, in the last three lines of the table. The new category NUM replaces the coding of the DELA, for which numerals are both DET and N. The PDET category is an additional reading for the forms in question. The PRES category is used for the forms *voici* and *voilà*, which are not in the DELA. In addition to these categories, the DM introduces a number of new features to distinguish some forms within a category, e.g. as seen in Table 4. All the features are documented in the dictionary documentation.

Extension of the DM

At the time of writing, the DM specifies a set of more than 538 000 word form-lemma pairs, with a total of more than 62 000 lemmas. For each category, Table 5 gives the number of entries, *i.e.* (*form, lemma, category*) triples, the number of forms and the number of lemmas[5].

Cat.	Entries	Forms	Lemmas	Description
A	56464	42404	14154	adjectives
ADV	2402	2389	2374	adverbs
N	83700	78135	36537	nouns
V	394014	298476	7672	verbs
INTJ	103	103	103	interjections
PFX	865	865	865	prefixes
CONJC	11	11	10	coordinating conjunctions
CONJS	66	66	33	subordinating conjunctions
DET	103	77	57	determiners
PREP	59	59	57	prepositions
PRO	144	125	109	pronouns
NUM	109	109	103	cardinal numbers
PDET	6	6	2	the predeterminers *tout, tous, toute, toutes,* and the adjectives *feu* and *feue* (e.g. *feue la reine*)
PRES	2	2	2	presentatives: *voici, voilà*

Table 5. Categories of the DM, with figures.

[5] Identical lemmas with two different categories are counted twice. Figures are obtained via a conversion of the NooJ generated dictionary of inflected forms into the *lexc* format of the XFST platform (Beesley and Karttunen 2003). In addition to the values given in the table, the DM contains a few functional compounds such as *duquel* or *à l'égard des*, to which correspond *sequences* of lemmas.

For the A, ADV, N, V and INTJ categories, values are approximately that of Table 2 (*int.* × *union* /100). The correspondence is not exact because we sometimes introduced some lemma distinctions not taken into account in Table 2, and a few minor modifications. For instance, the DM distinguishes explicitly two lemmas *mort_1* ("a dead person", 4 forms) and *mort_2* ("death", 2 forms).

Evaluation

To give a hint of what NooJ users working with the *delaf.nod* dictionary can expect using the DM, we present here a comparative evaluation on corpus, with two scenarios: morphological analysis and parsing. In both cases, the DM set of function words, including NUMs, is used with higher priority; evaluation is limited to lexical words only.

Morphological Analysis

Table 6 compares the results of morphological analysis on three texts: the *XML Language Specification*, a set of news articles from *Le Monde* and Jules Verne's *Le Tour du monde en 80 jours*[6]. The first three lines give the number of word forms, of word form types[7] and of *lexical* word form types, on which the evaluation focuses. Figures in the rest of the table are percentages of the lexical word form types (LWF).

91% to 92.8% of the LWF are analyzed the same way (either the LWF gets the same annotation or it is unknown to both dictionaries). The high percentage of unknown words in *XML* is due to the variable names in the XML grammar and XML examples. Otherwise, unknown forms are mostly proper names, plus some foreign words and spelling errors.

The number of word forms which are only known to the *delaf* is very small. These words are for the most part proper names, proper name adjectives, English words, words formed by affixation or composition[8]. Some of them have entered or re-entered French usage recently; as publication of the TLF went on in the 70-80s, Morphalou does not know

[6] Column *3 texts* gives the figures for the concatenation of the three texts. It is not the sum of the others as some forms appear in several texts.

[7] Word forms which are identical after uppercase-lowercase transformation are instances of the same word form type, except when the case difference yields a difference in the annotation.

[8] e.g. *Amérique, France, Bush, Henri...; grenoblois, brahmanique, irakien... ; basic, country, engineering... ; cofinance, prédéfinies, arabisation, délimiteurs...*

these words[9]. In this respect, one may note that the annotations are the most similar on the oldest text (87%).

	XML	Monde	Le Tour	3 texts
word forms	24513	9917	70877	105307
WF types	2823	3233	8908	12449
lexical WFT	2669	3061	8647	12167
same annotation	64.4	82.6	87	80.2
unknown to both	28.3	8.4	5.8	12.2
only in *delaf*	1.5	1.2	0.3	0.9
more in *delaf*	5.7	7.7	6.8	6.6
more in DM	0.1	0.1	0.2	0.1

Table 6. Morphological analysis comparison.

For 5.7 to 7.7% of the lexical word form types, the *delaf* produces more annotations than the DM. There are 825 extra annotations for 802 word forms. More than ¾ of these extra annotations may be analyzed as cases of conversion from one category to another, as illustrated in Table 7. The idea is that these forms have a fundamental category (*FC*), but may be used in the syntactic context of another (*Conv*), e.g. in *voir grand* ("see big"), *grand* becomes an adverb, in *état membre* ("member state"), *membre* becomes an adjective. The question whether these should be coded in the lexicon or handled by the parser is open.

word form samples	FC	Conv	#
bas, court, faux, fin, grand, grave, jaune, juste	A	ADV	36
grand, immobiliers, impatient, intégrale, littéraire, modéré, perdu, préférée, blindés...	A or V+PP	N	362
habitée, identifié, importé, nommé, prononcé...	V+PP	A	80
aboutissant, causant, combinant, déroulant...	V+G	A	17
membres, victime, nord, modèle, sable...	N	A	147
attention, ciel, grâce, paix, salut...	N	INTJ	11

Table 7. Examples of hard coded conversions in the *delaf*.

[9] e.g. computer science terms: *web, internet, formatage, implémentation,* or for historical reasons: *balkanisation, croate* (Morphalou only knows *serbo-croate*).

About 12% of the *delaf* extra annotations are additional verb readings, of which 80% are errors (e.g. *marchandises, hier, étranger, avis*) and the rest are the result of noun conversion (e.g. *candidates, paramètre*). The rest of the extra annotations is a mixture of *delaf* errors, debatable categorization choices (e.g. days of the week as ADV) and possible DM incompleteness. In any case, this incompleteness is very limited.

Finally, there are a few cases where the DM produces more annotations. These are verb forms: feminine or plural past participle of intransitive verbs or second person of impersonal verbs[10]. As mentioned above, The DM verb grammars overgenerate slightly.

This evaluation also revealed three function words missing in the DM (*ès*, *quiconque*, and *force* as a DET) and a couple of errors in the verb inflection grammars. These have been corrected. The DM is fairly reliable.

Influence on Parsing

The second evaluation scenario measures the influence of the dictionary change on parsing results. We use the non-deterministic chunker of (Trouilleux 2009a, 2009b) with the *delaf.nod* and the DM and compare results. The corpus is the *Le Monde* text; it contains 5297 chunks. As seen in Table 8, with the DM, the number of output chunks decreases by 227, while the number of correct chunks decreases only by 13. This yields a little loss in recall and a significant gain in precision.

	delaf.nod	DM	variation
output	7701	7474	–227
correct	5263	5250	–13
recall	99.36	99.11	–0.25
precision	68.34	70.24	+1.9

Table 8. Parsing results with *delaf.nod* and DM.

Considering that the parser has been designed with the *delaf*, these results suggest that the *delaf* extra annotations mostly introduce spurious ambiguities. Future work will consist in exploring further lexical category conversion in relation to parsing.

[10] e.g. *marchés, cheminée, souris...* ; *ventes, neiges, brumes.*

Conclusion

As evaluation shows, the DM is a fairly good coverage dictionary and a reliable substitute to the *delaf.nod* file. It opens up a world of opportunities for NooJ users: using constraints in grammars, adding information or entries, etc. The DM will be included in the NooJ distribution and will be freely downloadable from the author's web page[11]. We intend to maintain and further develop this dictionary, and would be happy to do so in a community framework. In this respect, integration of other NooJ resources will be considered, e.g. the dictionary of verbs of (Silberztein 2010). On a personal basis, category conversion and detection of words formed by affixation will be of prior interest for future work. Dictionaries for proper names and compounds would also be needed.

References

Beesley, K. R. and L. Karttunen 2003, *Finite State Morphology*. Stanford: CSLI Studies in Computational Linguistics.

Courtois, B. 1990, « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française* 87, 11, Paris: Larousse. http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html

Le Goffic, P. 1997, *Les Formes conjuguées du verbe français : oral et écrit.* Paris: Ophrys.

Mathieu-Colas, M. 2009, « Morfetik : une ressource lexicale pour le TAL », *Cahiers de lexicologie* 94, 137

Romary, L., S. Salmon-Alt and G. Francopoulo 2004, "Standards going concrete: from LMF to Morphalou", *Workshop on Electronic Dictionaries, Coling 2004*, Geneva. www.cnrtl.fr/lexiques/morphalou/ (ATILF/Nancy Université - CNRS)

Sagot, B. and K. Fort 2007, « Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire - Adverbes en *–ment* », *Actes du Colloque Lexique et Grammaire 2007*, Bonifacio, France

Sagot, B. 2010, "The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French", *7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malte. http://alpage.inria.fr/~sagot/lefff.html

Silberztein, M. 2003, *NooJ Manual*. http://www.nooj4nlp.net (220 pages, updated regularly).

[11] http://www.univ-bpclermont.fr/LABOS/lrl/spip.php?rubrique48

—, 2005, “NooJ's Dictionaries”, *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 »,. *Langages* n° 179-180, pp. 221-241. décembre 2010. Larousse.

Tolone, E. and B. Sagot 2009, “Using Lexicon-Grammar tables for French verbs in a large-coverage parser”, *Proceedings of LTC 2009*, Poznań, Pologne

Trouilleux, F. 2009a, Non Deterministic Chunking. *NooJ 2009 International Conference and Workshop*, Tozeur, Tunisia: Centre de publication universitaire.

Trouilleux, F. 2009b, « Un analyseur de surface non déterministe pour le français »*, Actes de la 16e Conférence sur le traitement automatique des langues (TALN’09),* Senlis, 24-26 juin 2009.