GéDériF: Automatic Generation and Analysis of Morphologically Constructed Lexical Resources

Fiammetta NAMER* Georgette DAL**

* « LANDISCO », Université de Nancy 2 - 23 Bd Albert 1er - BP 3397 - 54015 Nancy cedex. France
email : namer@clsh.univ-nancy2.fr.
** UMR 8528 « SILEX », CNRS & Université de Lille 3 - BP 149 - 59653 Villeneuve d'Ascq cedex. France
email : dal@univ-lille3.fr.

Abstract

One of the major frequent problems in text retrieval comes from large number of words encountered which are not listed in general language dictionaries. However, it is very often the case that these words are morphologically complex, and as such have a meaning which is predictable on the basis of their structure. Furthermore, such words typically belong to specialized language uses (e.g. scientific, philosophical or media technolects). Consequently, tools for listing and analysing such words can help enrich a terminological database. The purpose of this paper is to present a system that automatically generates morphologically complex lexical French items which are not listed in dictionaries, and that furthermore provides a structural and semantic analysis of these items.
The output of this system is a morphological database (currently in progress) which forms a powerful lexical resource. It will be very useful in Natural Language Processing (NLP) and in IR (Information Retrieval) applications. Indeed the system generates a potentially infinite set of complex (derived) lexical units (henceforth CLUs) automatically associated with a rich array of morpho-semantic features, and is thus capable of dealing morphologically complex structures which are unlisted in dictionaries.

Introduction

In text retrieval, we often encounter words that are not listed in general language dictionaries. For example, the adjective *détectable [detectable]* appears 45 times in (*Le Monde* 1993), hereafter *LM*, and in the (*Encyclopedia Universalis* 1995) hereafter *EU*. However, this adjective is not attested in the *Robert électronique* (*RE*), or the *Trésor de la langue française* (*TLF*), or the *Nouveau Petit Robert* (*NPR*), despite the fact that these three sources together cover the synchronic attested lexicon.

Very often, these words unlisted in dictionaries are complex lexical units (henceforth CLUs)[1] that, of themselves, have a meaning that can be calculated from their structure. For example, the meaning of the word *détectable* can be calculated from the meaning of the suffix *–able* applied to the verb *détect(er)*. *Détectable* indicates that the referent of the noun which governs it has the latent characteristic of being able to be detected[2], and that is exactly how the *EU* uses it in the following excerpt:

"l'anomalie reste [...] détectable par des méthodes biologiques." (*EU*, s.v. **hémoglobinopathies**)

*["the anomaly remains [...] detectable through biological methods." (EU, s.v. **hémoglobinopathies**)]*

For the most part, these constructed words come from specialized languages (scientific technolects, media technolects, etc.), and tools that allow them to be listed, analysed, or generated can be useful in enriching a terminological database.

The goal of this paper is to present an automatic generation and analysis system for CLUs which are *a priori* absent from general language dictionaries. The system, called GéDériF, is a product derived from the *MorTAL*[3] project. Between now and 2002, it will (semi-) automatically assign structures and semantic descriptions to approximately 15,000 attested French CLUs[4].

MorTAL analyses attested units. It is therefore inefficient for unlisted lexical units. However, it is apparent that some of the rules used in the automatic analysis of an attested lexicon can be used in automatic generation and analysis of units that are absent from the dictionaries.

Once we have looked at the progress status of the automatic processing of CLUs unlisted in dictionaries, we will go on to explain how we designed our generator-analyser, focusing on three construction operations for lexical units: *–ité*, *-able*, and *–is(er)* suffixation. We will round up this article with a short evaluation, and then formulate the conclusion.

Progress Status, or : Processing of CLUs not Listed in Dictionaries?

Analysers Based on Dictionaries

Most of the (rare) automatic systems that give information, no matter how minimal, about CLUs start with closed lexicons[5]. Such is the case, for example, of the French system developed by (Grabar N. & Zweigenbaum P. 1999). Its goal is to constitute a morphological database using the SNODEM medical termonology.

---

[1] (Froissart C. & Lallich-Boidin 1996) noted that 32% of the forms not recognized by their morphological analyser CRISTAL are constructed words (proper nouns make up the other large contingent of words unlisted in dictionaries and sigles).

[2] (See Dal & *al.* to appear).

[3] This project, which brings together Ch. Jacquemin, N. Hathout as well as the two authors of the present work, is funded by the Ministère de l'Education Nationale, de la Recherche et de la Technologie français [French National Ministry of Education, Research and Technology], as part of the program Actions Concertées Incitatives 1999 [Concerted Incitement Actions].

[4] This number corresponds to an estimation of the number of derivatives produced by the affixes *-(a)tion*, *-(at)eur*, *-able*, *-age*, *-aire*, *-al*, *dé-*, *-et(te)*, *-eux*, *-ifi(er)*, *-is(er)*, *-ité* and *-oir(e)*.

[5] In French, NLP attaches little attention to constructional information (Bouillon P. 1998: 48), which is considered less adapted to the field than inflectional information (Sproat R.W. 1992; Fradin B. 1994).

The system designed by (Savoy J. 1993) is another analyzer that uses an extensive lexicon. Unlike the former, it does a complete morphological analysis of untagged words of a text which is to be used in information retrieval (henceforth IR).

Whether they produce morphological resources or label texts, these various systems are at a loss when confronted with a term which is not in their original lexicon.

But not all analysis systems of (supposedly) constructed lexical units use a closed lexicon. This is especially true in the case stemmers, most of which form word families characterized by a common root and inflectional and derivational links. These processes can be broken down into two major groups: processes based on rules, which often make use of Porter's Algorithm; and processes that work solely or mostly in a statistical mode, such as the Automorphology program.

We will show below that although these approaches can process unlisted constructed words, they generate either noise or silence.

Porter's Algorithm

Porter's Algorithm (see (Porter M. 1980)) uses a rule-based suffix stripping system of whose application mode is a function (1) of the suffix[6] which is to be deleted (the concept of suffix is purely geographical in this case) and (2) of characteristics of the stem. For English, this algorithm is so effective in IR that the addition of linguistic features was considered unnecessary (no improvements in terms of results were observed, especially in the studies described in (Lennon M. & *al*. 1981)). However, for morphologically complex languages, such as Dutch, linguistic features do need to be integrated (Kraaij W & Pohlmann R. 1996.)

When used for French, the lack of linguistic features, and more specifically, the lack of lists of exceptions which affect the activation of the rules, leads to two types of errors. For example:

1. If we define the two suffix stripping rules as *–aille* and *-ite*, the analysis of the derivatives *ferraille [scrap iron]* and *ferrite [ferrite]* leads to the common initial sequence *ferr-*, which is used as a key for calculating the constructional family {*ferraille, ferrite*}. If we apply the same rules to the nouns *marmaille [a group of noisy children]* and *marmite [cooking-pot]*, the problem is clear. Because these two nouns have the same initial sequence *marm-*, they are put together as a constructional family.
2. On the other hand, the de-suffixing rules *–ement* and *-er* logically obtain morphological families such as {*gonflement, gonfler} [swelling, to swell]*, but do not recognize pairs like *achever [to complete]* and *achèvement [completion]*, which are related.

Moreover, unlisted CLUs are not exactly analysed, but simply integrated into families.

[6] Porter's Algorithm does not handle prefixes.

The Automorphology Program

Unlike the stemmer mentioned above, this program[7] works in a completely probabilistic way, taking suffixes as well as prefixes into consideration, and analysing texts on any subject and in any European language. Its goal is to offer a structural analysis of the units that it processes.

According to frequency criteria, affixes are learned by the system by matching identical sequences present in words which appear in the original text. As a result, we obtain a list in which each element is matched with its common sequence, followed by the family of endings grouped together. Thus, the analysis of *atomiser [to atomise]* and *atomique [atomic]* results in : *"atom"-"ique.iser"*.

The author claims that the results are satisfactory for texts of at least 100,000 words. However, a test conducted on a body of texts in French of over a million and a half words proved that many mistaken analyses remain (for example, *départ [departure]* and *département [department]* were analysed as having the common stem (*départ [depart]*), whereas some links which were to be expected, such as *région(s) [region(s)]* and *régional/aux [regional]*, were not made. These results confirm our belief that a constructional analyser of French language should include a linguistic component.

Summary: What our Approach Has to Offer

As a general rule, constructional analysers for the French language are either dependent on their original lexicons (therefore, unable to analyse units which do not appear in these lexicons), or yield limited results because of the lack of linguistic features. They generate insufficient or incorrect results. In addition, none of them, to our knowledge, offers a semantic analysis of constructed words, although such an analysis is a major component in NLP applications, as well as in text retrieval and information analysis.

Our goal is (1) to produce a lexicon of CLUs that do not appear in dictionaries, (2) to associate linguistically-motivated semantic and constructional information with the terms thus generated, like in the DériF system (see (Namer F. 1999), (Dal & al. 1999)), from which GéDérif, the system presented here, is derived, (3) to form micro-families. Our system takes most of its entries from *TLFnome*[8]. These entries are manually supplemented through a systematic check in *RE* and *NPR*. GéDériF's own output units have been added to the dictionary-attested entries.

In addition to the linguistic check that is done, GéDériF's most striking characteristic is that it gives its users a constructed words database that can be used lexically (through the word family) or morphosemantically (through the semantic relationship which exists between the constructed word and its primitive). The operation and results of GéDériF are outlined and illustrated in the §

[7] It can be downloaded at the URL: http://humanities.uchicago.edu/faculty/goldsmith/

[8] *TLFnome* is lexicon of inflected forms developed at the INaLF based on the nomenclature of the *Trésor de la Langue Française*. It currently contains 63,000 lemmas, 390,000 forms and 500,000 entries. It is in the course of being supplemented by 36,400 additional lemmas from the *TLF* index.

GéDériF System. First, we will linguistically legitimise the choice of generation and analysis in the § **Linguistic Legitimisation**.

The Constitution of the Generator-Analyser of Constructed Units Which Are not Listed in Dictionaries

As we mentioned above, our generator-analyser of constructed units that are *a priori* unlisted in language dictionaries is a system designed concurrently with the *MorTAL* project. It is therefore dependent on that project's progress. For this reason, it can only work currently on the suffixes *-able*, *-ité*, and *-is(er)*. We will concentrate our attention on these suffixes in what follows.

Linguistic Legitimisation

A system that automatically generates and analyses CLUs unlisted in dictionaries is, by its very nature, an over-generating mechanism. This over-generation needs to be linguistically verified. A generator that could produce a linguistic monster such as **infabricagiste* (see Gruaz C. *& al.* 1996) would be too powerful.

We have also been careful to insure that the results produced by GéDériF be linguistically motivated, both from a formal and a semantic point of view. We will illustrate this by analysing *–able, -ité* and *–is(er)* suffixations to the extent that it fits our hypothesis, and then we will produce all the possible combinations that these three operators authorize.

A Brief Overview of *-able, -is(er)* and *-ité* Suffixations

The French suffix *-able* forms only one categorical type of derivatives, i.e. adjectives. However, it can operate on two categorical base types, i.e. verbs ($mang(er)_V$ / $mangeable_A$ *[to* eat_V */* $edible_A$*]*) and nouns ($ministre_N$ / $ministrable_A$ *[*$minister_N$ */ potentially* $minister_A$*]*). The common semantic characteristic of all $Xable_A$*'s* is that they indicate that the referents of the nouns they modify have a capacity that can be revealed by a process (for more details on *-able*, (see Dal G. *& al.* 1999)).

The suffix *-is(er)* forms verbs and is also applied to two categorical base types: adjectives ($moderne_A$ / *modernis(er) [*$modern_A$ */ to modernize]*) and nouns ($bémol_N$ */ bémolis(er) [a flat musical* $note_N$ */ to put in a flat note, or to tone down a statement]*). Like its English homologue *-ize* (see Plag I. 1997), *-is(er)* covers a wide spectrum which is more or less obvious depending on the category of the base.

1. *Xis(er)'s* that are derived from adjectives express a condition change for the referents of their direct objects. In French, this characterization concerns all verbs derived from adjectives, no matter what operation formed them, because it is the only possible semantic relationship that can be established between an adjective and a derived verb.
2. *.Xis(er)'s* that are derived from nouns can express various processes depending on the meaning and/or the referential characteristics of the base noun:

- The base is a proper name referring to an individual with specific behaviour: *Xis(er)* describes the process of behaving in a similar manner as that particular individual (e.g. *socratiser [to Socratize]*).
- The base is a proper name referring to an individual who is well known for his or her work (literary, political, etc.): *Xiser(er)* describes the process of giving the direct object's referent a characteristic typical of the work in question (e.g. *brechtiser [to Brechtize]*).
- The base refers to an action or the protagonist of an action: *Xis(er)* describes the process of subjecting the direct object's referent to an action defined by the base or by one of its protagonists (*bémoliser; macadamiser [to macadamise]*). It can also be intransitive, and form a factitive (*pactiser [to make a pact]*).

Thus we see that the suffix *-is(er)* can be applied to semantic types with different bases, and seems to be the French verb-forming suffix *par excellence*.

Finally, the suffix *-ité* only forms one categorical type of derivatives (nouns), and operates on two categorical base types, adjectives ($absolu_A$ / $absolu ité_N$ *[*$absolute_A$ */* $absoluteness_N$ *]*), and less frequently, nouns ($édile_N$ / $édilité_N$ *[*$councillor_N$ */councillor's* $magistracy_N$ *]*). From a semantic point of view, *-ité* forms property nouns presented as objective (see Corbin D. to appear; Dal G. 1997).

Possible Combinations

We applied GéDériF to all the combinations that can be formed with the three suffixes described above. In this section, we will examine which combinations are *a priori* possible, and also which ones are in fact found in the attested lexicon in dictionaries. The latter will allow us to justify the GéDériF output.

[[Xis(er)] able] / *[[Xité] able]

According to the categorical characterisation given above, we see initially that the suffix *–able* can be applied to *-is(er)* verbs and to *–ité* nouns.

Furthermore, these configurations are semantically and syntaxically valid.

1. When applied to verbs, *-able* selects verbs that have at least one internal argument (direct object or locative argument: e.g. *skiable [skiable]*). However, most *-is(er)* verbs are transitive. Therefore, we have kept the structure [[Xis(er)] able].
2. On the other hand, *Xable's* derived from property nouns are very rare in the attested lexicon. The most recent one dates from the 16th century (see Dal G. *& al.* 1999; Dal G. *& al.* to appear). For this reason, we have rejected the [[Xité] able] structure[9].

****[[Xable] is(er)] / *[[Xité] is(er)]***

Derivatives using *–able* and *–ité* fulfil the categorical requirements of the suffix *–is(er)* in relation to the base that it selects (see § **Linguistic Legitimisation**).

[9] This decision is confirmed by the intuitively recognized outrageous character of sequences such as **absurditable*, **aciditable*.

From a semantic point of view, on the other hand, *-ité* nouns are rejected as bases because the semantic type does not correspond to any of the types accepted by *-is(er)*.

Applying the suffix *–is(er)* to *Xable's* also poses a semantic problem, though not as clearly. This is true at least in the case where the *Xable's* themselves are derived from verbs. The dictionaries attest only 14 *Xabiliser* verbs out of 700 *–is(er)* verbs. In addition, only 2 of those 14 derivatives (*navigabiliser [to make navigable]* and *respectabiliser [to make respectable]*) clearly have *–able* adjectives derived from French verbs as bases. This blockage can be explained as follows: *-able* adjectives express latent characteristics, that is to say characteristics which are endogenous to the referents of the nouns which govern them. However, in *Xiser's*, X_A describes the condition in which the entity finds itself after the process has taken place. It follows logically that X_A should express a characteristic that can result from a process taking place, therefore an exogenous process. Thus, the lack of *Xabiliser* forms in the attested lexicon is due to a semantic incompatibility between the *Xable's* and the requirements that *–is(er)* has for the adjectives that it selects.

As a result, we have not allowed GéDériF to apply the suffix *–is(er)* to *–able* adjectives derived from verbs. To confirm the validity of this decision, we conducted an automatic verification on the search engine www.yahoo.fr using 1287 *–abiliser* verbs created especially for the purposes of demonstration. Only 6 of these generated terms (approximately 0.5%) got positive results, and only half of those (*commutabiliser [commutabilize]*, *portabiliser [portabilize]*, and *variabiliser [valiabilize]*) had the structure that we had rejected. Consequently, the silence that we generate by refusing that structure is negligible compared to the noise that we would generate if we retained it.

[[Xable]ité] / *[[Xis(er)] ité]

As we have already mentioned, the suffix *–ité* can select adjectives and nouns. Outputs of *–is(er)* suffixation are thus immediately excluded as bases.

Products of *–able* suffixation are both categorically and semantically licit bases. They are categorically licit because they are adjectives, and semantically licit because there is compatibility between the meaning of *–able* adjectives and *–ité* suffixation, which requires bases expressing objective properties. Any *–able* adjective can thus *a priori* result in an *–ité* property noun[10].

Before authorizing GéDériF to automatically apply the suffix *–ité* to all *Xable's*, we examined the special case of *–isable* adjectives.

If what we have just put forward is true, *Xisable's*, which are merely a special case within *Xable*, should be able to be used as *–ité* suffixation inputs. Surprisingly, we see that out of 214 *-abilité* property nouns taken from the dictionaries, only two (*hypnotisabilité [hypnotisability]* and *polarisabilité [polarisability]*) have an *–able* adjective with *–is(er)* suffix in its structure as a base (although the dictionaries offer approximately 70 possible bases, e.g. *commercialisable [marketable]*). In order to determine whether we should keep this configuration or not, we sought an explanation for the absence of *–isabilité* nouns in the dictionaries.

The only reason that we found for that absence is related to performance. We hypothesize that the small proportion of *Xisabilité* derivative forms in general language dictionaries is due to the fact that these derivatives include three successive constructional operations, and even a fourth one when the adjectival base of the *–is(er)* verb is also constructed. These numerous operators can cause a problem for semantic calculations, which make alternative strategies preferable. Rather than " La commercialisabilité de ce produit fera l'objet d'une étude de marché " *["The marketability of this product will be the subject of a market study,"]*, we might prefer "La possibilité de commercialiser ce produit […]" *["The possibility of marketing this product […] "]*. However, since the reason for the blockage is not specifically linguistic, it seemed justifiable to also generate *Xisabilité's* (see § **Quantitative Assessment**).

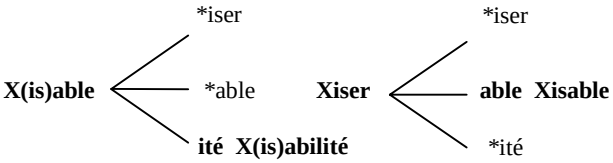
The GéDériF System

There are two components to the GéDériF system. It starts with a lexicon of approximately 70,000 non-inflected forms tagged with Brill's tagger (see (Brill E. 1980) and (Lecomte J. & Paroubek P. 1996)). The first component automatically generates new lexical entries, according to a certain number of criteria (see § **Generating the Lexicon of Possible Words**). The second (see § **How the DériF Analyser Works**) makes a constructional analysis of the words in the new lexicon, and produces a lexicon with additional constructional and semantic information as output.

Generating the Lexicon of Possible Words

The generator produces lexical units by affixing existing lexical units. This task presupposes that the base units are semantically compatible with the affixation operator. Given the current lack of semantic features in the monoconstituated lexical entries, the generator can take only constructed words as the base for the units that it will generate: indeed, the meaning of such constructed words can be calculated according to the semantic instruction of the affix that produced them.

In the case that we are dealing with here, GéDériF selects *–able* and *–iser* words from the lexicon, and concatenates all of the licit *–able*, *-iser* and *–ité* possibilities. Thus, out of all the combinations that are *a priori* possible for *Xiser* and *Xable* forms of lexical units, only the ones in the boxes below are valid (see § **Linguistic Legitimisation**):



[10] The number of *–abilité* derivatives attested in our referent dictionaries (the *TLF*, the *RE* and the *NPR*) confirms this. 214 of the 1409 *–ité* nouns that we compiled (=15.2%) are derived from *–able* adjectives.

The only forms retained from the results obtained are those absent from the original lexicon.

Using *Xable* and *Xiser* lexical units together with the suffixes *–able* and *-ité*, GéDériF produced 2691 constructed words which were absent from the original lexicon, and which we consider linguistically possible.

How the DériF Analyser Works

The second component of GéDériF is the constructional analysis module, called DériF. It acts upon the constructed words appearing in the original lexicon which has been supplemented by the terms obtained through automatic generation. The result of the analysis of a word W is a triplet that includes: (1) the morphological **parse tree** of W in brackets, (2) all of the lemmas that belong to W's **morphological family**, and (3) the **meaning constructed** by the last affix that operated to produce W, given in the form of a semantic relationship between W and its base.

Thus for W = *localisabilité [localizability]*, the result is as follows:

(1)	[[[[local ADJ] iser VB] able ADJ] ité NOM]
(2)	(localisabilité, localisable, localiser, local)
(3)	**:: Propriété de ce qui est localisable** *[Property of that which is localizable]*

Table1

Since DériF has already been described in another work (Namer F. 1999), it is presented only briefly here.

The driver examines the entry word W and, if necessary, calls up the function F_S performing the analysis of the S suffix of W. First, F_S checks whether prefixes applicable to W's suffixed base exist. Then F_S truncates W according to its suffix S, and goes on to do semantic and categorical calculations, and to pair the allomorphs. It then sends the results R to the driver. The driver repeats its examination of R until it arrives at a primitive, then displays the results. The most important part of the task is performed by the F_S functions, which we will show below.

***Xité* and *Xiser* Analysis**

We will limit ourselves here to a brief presentation of the algorithm underlying the functions $F_{ité}$ and F_{iser} which analyse respectively the lexical units of *Xité* and *Xiser* forms.

-Ité and *–is(er)* can be applied to adjectival bases with similar semantic characteristics (this is based on the fact that, according to the dictionaries, the same adjective can sometimes be used as input for the two suffixations, as is the case for *absolu [absolute]* and *adverbial [adverbial]*). These bases are also subject to similar allomorphic variations. Thus the functions $F_{ité}$ and F_{iser} share a large part of the pairing system for allomorphic variations of the base X resulting from the troncation of *–ité* or *–iser*, according to X's final sequence (with a few exceptions). Thus the pairing rule **ar|ier** leads to the allomorph *régulier*$_{ADJ}$ *[regular*$_{ADJ}$*]* from the base *regular-* which also appears in *régularité*$_{NOM}$ *[regularity]* and *régulariser*$_{VERB}$ *[to regularize]*. The base *human-* uses the rule **an|ain** to pair with *humain*$_{AJD}$ *[human]* in *humaniser*$_{VERB}$ *[to humanize]* and *humanité*$_{NOM}$ *[humanity]*.

Other variations can bring to light an infralexical unit (ILU) which identifies a foreign base, noted as FWD (e.g. *virgin-* in *virginité [virginity]* or *virginiser [virginize]*). The program picks it out from a special table that lists all bases that are not found in the referential and that come from Latin, Greek, German, etc., along with their translations, approximative when necessary, as well as the grammatical categorization of the translations (e.g. *virgin-* is translated *vierge* ADJ).

In the results, the ILU is retained in element (1) of the triplet, while its translation appears in elements (2) and (3) (see Table 1).

Results

We will conclude this brief presentation of GéDériF with a description of the various kinds of results obtained.

Parse tree

In DériF, the parse tree for the units described is structured with brackets and tags. For example:

(a) recristalliser *[to recrystallize]* =>[re [[cristal NOM] is(er) VB] VB] (recristalliser, cristalliser, cristal)
(b) inaliénabilité *[inalienability]* => [[in [[aliéner VB] able ADJ] ADJ] ité NOM] (inaliénabilité, inaliénable, aliénable, aliéner)
(c) biodégradabilité *[biodegradability]* => [[[bio NOM] [[dégrader VB]able ADJ]ADJ] ité NOM] (biodégradabilité, biodégradable, dégradable, dégrader)

A representation of this type shows the various consequences of the constructional operators, seen in the morphological family formed concurrently (and displayed as a parenthesized list).

Thus the bracketed structures (a) and (b) show the relative order of the suffixation and prefixation operations, namely suffixation FOLLOWED BY prefixation for *recristalliser [to recrystallize]* (the bracketed diagram accounts for the fact that the prefix *re-* is applied to the suffixed verb *cristalliser [to crystallize]*, and therefore constitutes the most peripheral prefix); prefixation FOLLOWED BY suffixation for *inalienabilité [inalienability]* (in this derivative, the suffix *–ité* is applied to the prefixed base *inalienable*).

Finally, example (c) concerning *biodégradabilité* illustrates a suffixation operation (by *–ité*) on the compound adjective (*biodegradable)* resulting from a compositional operation applied to a suffixed adjective: the infralexical noun *bi(o)-* being compound to the adjective *degradable*.

Gloss

In the linguistic model underlying DériF[11], the CLU's are given metalinguistic definitions (see especially (Corbin D. 1993)). The glosses that are automatically assigned to the DériF entries are deliberately formulated in natural language so that they can be used in LP and IR.

In DériF, the glosses show the most peripheral constructional operation corresponding to the semantic characterization described in the § **Linguistic**

[11] A constructional morphology model developed in France in the UMR SILEX under the guidance of D. Corbin.

Legitimisation. Although the gloss only reflects the last constructional operation, semantic information from previous levels, if there are any, can be recovered and used.

Let us look at the case of *localisabilité [localizability]*, which is a $[X_{ADJ}$-ité$]_{NOM}$ type of lexical unit. Its corresponding gloss therefore is an instance of the generic gloss: **property of that which is X_{ADJ}**, associated with nouns produced by applying *–ité* to adjectives. The complete analysis of *localisabilité* is given in Table 1, and the value of the gloss is recalled as follows:

localisabilité ==> :: Propriété de ce qui est localisable

Localisable is a $[X_{VB}$-able$]_{ADJ}$ type lexical unit. This type of derivative expresses the possibility for the referent of the noun that governs it to have the process expressed by *X* applied to it. This means that a first approximation of the gloss is **that which can be X_{VB}**. The semantic analysis of *localizable* is therefore:

localisable/ADJ : […] (…) **:: Que l'on peut localiser**
[Which can be localized]

Finally, the gloss that corresponds to the analysis of a $[X_{ADJ}$-iser$]_{VB}$ type lexical unit expresses in natural language the semantic instruction associated with the suffix *–iser* when it is applied to adjectives. One way of rendering the constructed meaning of the derivative is **to make X_{ADJ}**. Applied to *localiser*, this gloss gives:

*Localiser/*VB : […], (…), **:: Rendre *local***
[To make local]

By moving from one related form to another, the meaning of the noun *localisabilité* can be reconstructed from its primitive *local*. This can be represented in the following semi-formal notation:

meaning_of(**localisabilité**)=propriété_de_ce_(que_l_on_peut(rendre(**local**)))
[property of that which can be made local]]

Therefore, the GéDériF system is not simply a generator of CLUs *a priori* unlisted in general language dictionaries. It also automatically assigns a tagged structure and a gloss to each unit it generates.

Assessment of the Results

We conducted a double series of quantitative and qualitative tests to assess the results presented above. First we tried to assess the proportion of invented terms that actually appear in documents. This calculation was done using 2691 terms constructed by means of licit combinations of the suffixes *–ité*, *-able* and *–iser*. We then manually verified the validity of the formal, structural and semantic analyses produced automatically by the GéDériF analyser on all of the lexical units.

Quantitative Assessment

Two Types of Searches

In order to validate the linguistically pre-filtered lexicon (see § **Possible Combinations**) obtained from the GéDériF generator (see § **Generating the Lexicon of Possible Words**), we conducted two series of verifications which allowed us to establish what proportion of these 2691 terms actually appears in documents.

First, we systematically verified the presence of the elements of this lexicon in the *Encyclopedia Universalis* and in the terminology review *La Banque des Mots*, which draws from a variety of sources (scientific and economic reviews, etc.). These two resources were chosen because they are representative of different fields, and therefore reveal various technolects. In order to insure that the *EU* and the *BDM* are favourable to the emerging of constructed terms unlisted in dictionaries, we conducted an automatic verification as a control on two corpora of texts, 8M each, containing (respectively) articles from the newspaper *Le Monde* from the year 1992, and bibliographical notices from the food industry taken from the database PASCAL[12].

Concurrently, we made a program that automatically checked the search engine www.yahoo.fr for each of the terms generated. The response to these requests specified how many occurrences, if any, were found. This allowed us, to a certain extent, to weigh the validity of each term.

Results

The results of the searches done on the corpora from *Le Monde* and the food industry were practically nil (0.9% success out of 2691 terms), which confirms that (1) the words tested were too specialized in nature for a journalistic corpus, (2) the meaning of the words covers a range of specializations too wide for their presence to be notable in a corpus specialized in a single field.

As for the results of the searches done on *EU*, the *BDM* and on the Web, they are recorded in Table 2.

		Xisable	***Xisabilité***	***Xabilité***
Quantity (total=2691)		711	833	1103
Experiment1: Number **of occurrences in EU and BDM**		39	2	56
Experiment2: **occurrences on the Web**	**Success :** Nb/ **percentage**	101/ **13,4**	18/ **2,1**	232/ **21**
	Nb **fewer than 10 occs.**	94	15	197
	Nb **more than 10 occs.**	7	3	35
Overall occurences (the Web+EU+BDM) Nb **/percentage**		12 / **14,8**	18 / **2,1**	246/ **22,3**

Table2

The first two raws describe the spread of terms generated according to suffix combinations. The third raw indicates how many of these terms were found through manual search in *EU* or the *BDM* (validation *Experiment 1*). The three following raws summarize the positive results

[12] Scientific text database developed at and maintained by the INIST-CNRS.

obtained on the Web (validation *Experiment 2*). We broke these results down (raws 5 and 6) according to the number of occurrences reported by www.yahoo.fr. Finally, the last raw sums up the results of the two experiments (identical results, of course, being counted only once). These results call for some additional explanations:

1. The small percentage of *Xisabilité* terms (2.1%) used in the corpora confirms the hypothesis that a construction that combines more than two suffixes can cause performance problems, and that common usage tends to adopt alternative strategies to avoid forming such terms (see § **[[Xable]ité] / *[[Xis(er)] ité]**).
2. However, we get good results from the other two types of constructions. The arbitrary distinction between "more than 10 occurrences" and "less than 10 occurrences" in *Experiment 2* is also a possible indication of the multiplicity of the fields which generate the largest percentage of terms. A more exhaustive search would undoubtedly bring to light the fields in which the most terms develop. Such an experiment is outside the scope of our work, our goal being simply to show that such unlisted words are used.
3. Finally, the comparison between the results obtained through *Experiments 1* and *2* confirms that the lack of results on the Web does not prove that the term is not used, as is attested by *théâtralisable [theatralizable]* and *interdéfinissabilité [interdefinabiliy]* (*EU*), or *égouttabilité [drainability]* and *pluralisable [pluralizable]* (*BDM*), for example. We can therefore assume that the percentages obtained indicate a **minimal** number of terms used compared to those generated. In fact, among the terms automatically generated, no matter what licit combination of suffixes is used (*-iser, -able* and *–ité* being only one illustration of our system), we can expect **at least** 15-20% of words already used in one or more specialized fields.

What remains to be determined is whether GéDériF can give an appropriate constructional and semantic analysis to these terms whose automatic generation is justified by their proven or probable usage. Such analyses are *sine qua non* for being utilized in a database by an IR ou NLP user. This is precisely what we will present in the following paragraph.

Qualitative Assessment

Our second series of tests concerned the evaluation of the quality of the lexicon that was generated, from three points of view – formal, structural and semantic.

Before generating our lexicon, we defined a certain number of linguistic safeguards (see § **Possible Combinations**). In addition, the lexicon that was produced encountered no categorical problems (only categorically licit lexical units were generated), and *a priori* very few formal allomorphic problems. The risk of error is thus minimalized from the start.

In fact, the results that were obtained automatically are generally good, although in some cases they could be improved (or were erroneous). Indeed, the automatically generated and analysed lexicon inherited imperfections from the analyses already implemented on the attested lexical units.

Thus, DériF does not presently process cases of structural ambiguity. Among the suffixes that have been studied to date, such cases of structural ambiguity concern mostly derivatives that include *–able* in their structure. A case of particularly ambiguous structure is presented by adjectives of the *inXable* form when *inX* and *Xable* are respectively a verb and an adjective which are attested or possible, e.g. *inversable [reversible / unpourable]* derivable from *invers(er) [to reverse]* or from *versable [pourable]*[13]. DériF is currently programmed to regard all *inY* forms of adjectives as antonyms of Y (because it is most frequently the case). The *inXable(s)* in question are therefore given only a ([in[X ADJ]ADJ]) structure and a ("non-X") gloss. Naturally, this programming has repercussions on property nouns with *–ité* correspondants. However, this imperfection can be ignored, since, in any case, it only generates silence.

The other problem to be pointed out concerns *Xisable / Xisabilité* lexical units, when *X* = a proper name (e.g. *pantagruélisable / pantagruélisabilité [Pantagruelizable / Pantagruelizability]*). The problem with these sequences, discussed briefly in § **Linguistic Legitimisation**, is that *-is(er)* can be applied to proper names. But, according to the semantic content of the propoer noun, the *–is(er)* verb can be intransitive or transitive. Only in the latter case can it become the base of an *–able* adjective derivative (and through transitivity, an *–abilité* noun derivative). Since the DériF entries do not include semantic information (see § **The GéDériF System**), the GéDériF results suffer from this lack of information. In the current state of affaires we cannot automatically distinguish between units such as *brechtisable / brechtisabilité [Brechtizable / Brechtizability]* which are possible, given the substance of the proper name *Brecht*; and *pantagruélisable / pantagruélisabilité*, which are unlikely, given the content of Pantagruel. We have chosen to continue to automatically produce these two types of derivative structures for two reasons. First of all, because the base referent for the *Xiser* with *X* = proper noun, is an individual who is more often known for his or her work than for his or her specific behavior (for simple pragmatic reasons). Secondly, because the semantic calculation of an adjective like *pantagruélisable* poses more of a referential problem than a strictly linguistic one.

Conclusion

The presentation that we have made here concerning the possible combinations of the suffixes *–able*, *-ité* and *-is(er)* could be easily applied to other combinations widely used in technolects, and nonetheless unlisted in dictionaries. For example, the process that was followed here could also apply to the following combinations: **(i)** *-el + -is(er)* : *fictionnaliser* (*LM*), **(ii)** *-is(er) + -ation* : *ethnicisation* (*ibid.*), which could conceivably be combined with operation (i) (*fictionnalisation* (*BDM*)), **(iii)** *–ation + -el* (*civilisationnel* (*LM*)), which could conceivably be combined with operation (ii)

[13] Other examples: importable [importable / uncarriable], invalidable [unvalidatable / invalidable].

(*organisationnel* (*ibid.*)), **(iv)** *–if(er)* + *-ation* (*taudification* (*ibid.*)), etc.

Eventually, GéDériF will be able to generate, analyse and gloss a potentially infinite number of linguistically verified CLUs, by simply repeating the operations on the system outputs.

References

BDM = La banque des mots, revue de terminologie française publiée par le conseil international de la langue française, Paris, Conseil international de la langue française.

Bouillon, Pierrette (1998) : *Traitement automatique des langues naturelles*, Paris-Bruxelles, Duculot.

Brill, Eric (1992) : "A simple rule-based part of speech tagger*", in Proceedings of the 3d Conference ANLP-ACL*, Trento.

Corbin, Danielle (1993) : "Morphologie et lexicographie : la représentation du sens dans le *Dictionnaire dérivationnel du français*", *in* Hulk A., Melka F. & Schroten J. éds, 63-86.

Corbin, Danielle (to appear) : *Le lexique construit*, Paris, Armand Colin.

Dal, Georgette (1997) : "Du principe d'unicité catégorielle au principe d'unicité sémantique : incidence sur la formalisation du lexique construit morphologiquement", *in* P.-A. Buvet, S. Cardey, P. Greenfield & H. Madec éds, *Actes du colloque international Fractal'97, BULAG* numéro spécial, 105-115.

Dal, Georgette, Hathout, Nabil & Namer, Fiammetta (1999) : "Construire un lexique dérivationnel : théorie et réalisations", *in Actes de la VI^e^ conférence sur le Traitement Automatique des Langues Naturelles (TALN'99), Institut d'Etudes Scientifiques de Cargèse*, Corse, 12 -17 juillet 1999, 115-124.

Dal, Georgette, Hathout, Nabil & Namer, Fiammetta (to appear) : "Une base de données constructionnelles expérimentale : le projet MorTAL", *in* P. Boucher ed., *Morphology book*, Cambridge Mass., Cascadilla Press.

Dal, Georgette & Namer, Fiammetta (sous presse) : "Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations", *in* C. Jacquemin éd., T.A.L..

EU = CD-ROM *Encyclopaedia Universalis*, version 2.0, Paris, Encyclopaedia Universalis, 1995.

Fradin, Bernard (1994) : "L'approche à deux niveaux en morphologie computationnelle et les développements récents en morphologie", *T.A.L.* 35/2, 9-48.

Froissart, Christel & Lallich-Boidin Geneviève (1996) : "Morphologie robuste et analyse automatique de la langue : étude réalisée à partir des corpus de l'évaluation GRACE", *in Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, 1996, 88-96.

Grabar, Natalia & Zweigenbaum, Pierre (1999) : "Acquisition automatique de connaissances morphologiques sur le vocabulaire médical", *in Actes de la VI^e^ conférence sur le Traitement Automatique des Langues Naturelles (TALN'99), Institut d'Etudes Scientifiques de Cargèse, Corse, 12 -17 juillet 1999*, 175-184.

Gruaz, Claude, Jacquemin, Christian & Tzoukerman, Evelyne (1996) : "Une approche à deux niveaux de la morphologie dérivationnelle du français", in *Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, CLIPS-IMAG, les 13 et 14 nov. 1996, 107-114.

Kraaij, Wessel & Pohlmann, Renée (1996) : "Viewing stemming as recall enhancement", in *Proceedings of ACM-SIGIR 96, Conference on Research and Development in Information Retrieval*, 40-48.

Lecomte, Josette & Paroubek, Patrick (1996) : "Le catégorisateur d'Eric Brill. Mise en œuvre de la version entraînée à l'INaLF", rapport technique, Nancy, INaLF-CNRS.

Lennon, M. ; Pierce, D. ; Tarry, B & Willett, P. (1981) : "An evaluation of some conflation algorithms for information retrieval", in *Journal of Information Science*, n° 3, 177-183.

LM = CD-ROM du journal *Le Monde* pour 1992, Le Monde / Research Publications International, 1993.

Namer, Fiammetta (1999) : "Le traitement automatique des mots dérivés : le cas des noms et adjectifs en *-et(te)*", *in* D. Corbin, G. Dal, B. Fradin, B. Habert., F. Kerleroux, M. Plénat & M. Roché éds, *La morphologie des dérivés évaluatifs*, *Silexicales* 2, Villeneuve d'Ascq 169-179.

NPR = Le Petit Robert. Dictionnaire de la langue française. Version électronique du *Nouveau Petit Robert*. Disque optique compact CD-ROM, Paris, Dictionnaires Le Robert / van Dijk, 1996.

Plag, Ingo (1998) : "The polysemy of *–ize* derivatives : on the role of semantics in word formation", *in* G. Booij & J. Van Marle eds, *Yearbook of Morphology 1997*, 219-242.

Porter, Martin (1980) : "An algorithm for suffix stripping", in *Program*, n°14, 130-137.

RE = Le Robert électronique DMW, Disque optique compact CD-ROM, Paris, Dictionnaires Le Robert, 1994.

Savoy, Jacques (1993) : "Stemming of French Words Based on Grammatical Categories", *JASIS: Journal of the American Society for Information Sciences*, vol. 44 : 1, 1-9.

Sproat, Richard William (1992) : *Morphology and Computation*, Cambridge, Massachusetts / London, England, The MIT Press.

TLF = Trésor de la langue française. Dictionnaire de la langue du 19^e^ et du 20^e^ siècle (1789-1960), 16 vol., Paris, Éditions du CNRS (t. 1-10) / Gallimard (since vol. 11), 1971-1994.