

Abstract

This paper presents a number of recent studies in French morphology which make a massive use of data. These data (derived words) have been automatically collected from digital corpora, mostly from the Web. The main point developed here is that this massive increase in the amount of available data can substantially modify the results of a morphological study, and can lead to new theoretical conclusions that would not have been possible with traditional data such as wordlists gathered from dictionaries. However, using the Web as a corpus brings up several technical and methodological questions, which are dealt with through examples and discussions about the different tools and techniques available.

Extensive data for morphology: using the World Wide Web

Nabil Hathout, Fabio Montermini and Ludovic Tanguy

CLLE-ERSS (UMR 5263) Université de Toulouse & CNRS

1 Introduction

Morphology, as the other fields of linguistics, benefits greatly from using corpora. By nature, it is perhaps the most able to deal with large amount of electronic data, due to the ease with which word forms are gathered and processed by simple computer programs. This is why morphology is a linguistic domain whose evolution is directly connected to the amount of available data, and it is understandable that its evolution partially paralleled the technological evolution in data retrieval, from paper word lists and dictionaries to text corpora and, lastly, to the World Wide Web.

A question which is often raised in corpus linguistics concerns the quantity vs. quality issue: more data may correspond to a less controlled corpus, and an increase in the amount of available data leads to an increase of noise, but also of relevant data. However, this question can be considered in a different way from a morphologist's point of view. Since it focuses on the low-level mechanics of word coining (in our case, mostly suffixation), whose interactions with other linguistic phenomena are somewhat reduced, the study of (derivational) morphology can exploit raw material that could be considered unsuitable for other studies. On the contrary, more data is indeed

better data. In fact, only very large corpora may allow the observation and the study of rare phenomena, which may be too little represented, if at all, in a traditional corpus (such as a list of dictionary entries). Data-intensive morphology can also lead to new perspectives on previously studied mechanisms, and easily confirm or invalidate intuition and conjectures.

This paper will first describe the techniques used for gathering word forms from the Web, providing researchers with quantities of new data, sometimes up to ten times more than with previous methods. We will then present a number of recent studies that have used such extensive data focusing on specific French suffixes (*-esque*, *-este*, *-able*, *-ment*). Each of these case studies will show how the use of greater quantities of data led to new insights on a given morphological phenomenon.

2 Collecting morphological data on the Web

In the past, the data used for most morphological studies consisted of word lists manually extracted from dictionaries or corpora. Increased access to digital resources has, however, made it possible to easily automate the gathering of such data: electronic dictionaries such as the TLFi (*Trésor de la langue française informatisé*) or text databases such as Frantext can provide large lists of words containing a given string of characters (corresponding e.g. to a suffix, such as *-able*) in a matter of seconds. Such a simple operation would have taken several months some fifteen years ago.

In addition to traditional digital corpora (mostly consisting, for French, of literary texts and newspaper archives), there is an increasing use of the World Wide Web as a resource for linguistics studies. Access to this

resource can easily be obtained through generic all-purpose search engines such as Google or Yahoo, which allow researchers to instantly get contexts and frequency of use for a given word form.

Given the sheer size of the data searched, the Web can be used to extend the study of a particular scheme, giving access to many more occurrences of a target form, and providing a more intricate insight of its meaning(s) and use. Even more interestingly, the Web can also be used to discover word forms which have not been recorded in lexicographic works (due to their rareness and/or novelty), and whose very existence would have remained a simple conjecture. For the latter type of research, the Web is superior to all existing traditional corpora because of its size, variety and constant evolution. The Web has thus been successfully used to discover words previously considered to be unlikely or even theoretically impossible to coin. For example, one can find the prefix *anti-* attached to simple (non-constructed) adjectives, such as in *anti-triste* ('anti-sad') or *anti-obèse* ('anti-obese'), and even to adjectives following the *V-able* scheme, such as *anti-inflammable* ('non-flammable'). This latter case was previously declared as theoretically impossible for instance in Fradin (1997).

Using the Web for linguistic purposes, however, is subject to both methodological and technical questions.

2.1 Using the Web as a corpus for linguistics

Of course, the Web cannot be considered as just a huge data store which linguists can use without taking any precaution.

The advocates of the Web as a corpus emphasise its many advantages:

its size, the presence of many different text types (including some that cannot be found anywhere else), the variety of languages used, its constant evolution, etc. According to them, every corpus study should use the Web as soon and as widely as possible. In fact, many researchers have been doing so for many different tasks, in domains ranging from translation to collocations. However, most of the works of this kind use the Web with little questioning of its content and are mainly related to computational linguistics.

The opposite point of view is to consider that the Web is not a corpus at all. The following arguments have been used in this direction (Lüdeling, Evert and Baroni, 2007):

- it has not been constructed or balanced in any way (as compared to generic corpora such as the British National Corpus) and cannot be viewed as representative of anything except itself;
- one cannot find the most simple information about it as a whole (the size of the Web cannot be measured, and only a part of it is accessible through search engines);
- most Web pages provide no information regarding their author (age, sex, nationality, date of writing, or even proficiency in the language used);
- the Web is constantly evolving (pages appear and disappear), making most experiments impossible to reproduce;
- only the crudest methods of access are available (restricted to simple keywords search).

It is then understandable that some linguists simply refuse to consider data from the Web as linguistic material useful for research. In a less drastic

way, many researchers are simply using the Web as a source for building traditional corpora, after a careful selection of documents. A great amount of work is also in progress for dealing with some of the above questions, for example efforts are made for archiving Web pages¹, identifying Web text types (Santini, 2006), and even building a search engine able to provide sophisticated modes of access to Web pages (Resnik and Elkiss, 2005). However, none of these improvements can today compete with the generic search engines. Their speed, ease of use and amount of data available through make them the major access point to Web data, despite their many limitations.

All these problems cannot be solved, and as a result each occurrence has to be manually checked.

2.2 Using Web search engines

Any linguist can easily use a search engine to instantly verify her/his own intuition, or check some uses of a word, without needing any specific software or computing skill: the only thing to do is type a word form on the home page of the search engine. The given results are the approximate number of Web pages known by the engine to contain the word, and the list of Web addresses of these documents illustrated with short text extracts.

Doing this on a larger scale, however, calls for specific programs that can automate the querying of search engines. This can be done on several levels. Tools such as Webcorp (www.webcorp.org.uk) are built on top of search engines for linguistic purposes. Webcorp provides concordances and collocates, which are very useful for syntax, semantics or discourse analysis.

¹ A good example is given by the “Way Back Machine” at <http://www.archive.org>

But, morphological studies need to have access to subparts of words: the search for forms in this case cannot be done without technological means of handling substrings of graphic words. The basic tool needed should be able to retrieve word patterns, such as those commonly expressed by wildcards, e.g. “*able” which stands for “any word ending with *able*”. Search engines do not provide this, so morphologists can greatly benefit from dedicated software such as Walim (Namer, 2003). This tool uses computational morphology techniques to automatically generate inflected word forms, given a list of bases and a derivation scheme (i.e. a suffix). It then uses the word forms as queries to a search engine, and only retains those for which at least one occurrence can be found.

Using the same principle, and adding some features for the discovery of new word forms, Webaffix (Hathout and Tanguy, 2002) can be used in two different ways to gather derived words from the Web.

The first method, based on hypothesis testing, follows the same principle as Walim. The main difference between these tools is that Webaffix places emphasis on the filtering of the raw results from the search engine: it checks for the correct target language, typos, bad word cuttings, etc. It can also perform a more restrictive selection, in retaining only derived words that co-occur with their base forms in a Web page, an efficient criterion for the morphological link between the two. For instance, *copolymérisable* (‘copolymerizable’) will only be retained as a legitimate *-able* adjective form if the deduced base verb *copolymériser* (in any of its inflected forms) appears in its vicinity.

The second method takes advantage of some Web engines that allow

limited use of wild cards in the queries (seeabove),² and is thus able to gather new word forms in an inductive way, without any assumption or knowledge of the base form. An automatic analysis is performed on each new word in order to calculate its base form, and the cocurrence test is performed to ensure a correct analysis.

This second method, which make no assumption regarding the base form, has been thoroughly used for instance in (Hathout, Plénat and Tanguy, 2003) to gather many new *-able* words whose base cannot be found in any major dictionary (e.g. acronyms or proper names).

2.3 Caveats

One very common use of Web engines by linguists is the search for frequencies. In morphology, frequencies are useful for measuring the productivity of a given process, and for comparing two concurrent word forms (see Baayen, 1991, Fradin *et al.*, to appear). Such experiments, if performed on the Web, mostly rely on the comparative numbers of occurrence of different word-forms, as indicated by any search engine for a given query. The use of these raw numbers must be subject to a lot of caution for several reasons. First, it indicates a number of documents, not words, and several occurrences of the same word form in the same Web page only count as one. The other obstacle, more related to the nature of the Web itself, is the duplication of occurrences, which covers different phenomena such as quotes, plagiarism, or simply the verbosity of a given Web page author which can lead him to use a newly coined words dozens of time. The last problem that

² Unfortunately, this method is no more operational. It was relying on the Altavista search engine which stopped accepting wild cards in 2003.

these numbers rise comes from search engines themselves. These systems are conceived to deal with huge quantities of data and to give an instant answer to the user. This often leads them to produce approximate and unstable numbers (the same query can give different results even within a few hours). In addition, the commercial concurrence between the major search engines often makes them boast more pages for a given query than they can really display. All these facts should lead to increased carefulness when using these numbers in a quantitative analysis, such as the comparison of two word-forms' frequencies. If the difference between presence and absence can be taken as significant, as well as very large differences between two frequencies (such as 2 vs 100,000 documents), smaller quantitative differences (such as 10 vs 15 documents) should not.

Other limitations are due to the technology used by search engines for automatic language identification: many resulting documents for a given query are not written in the selected language, and should not be considered as valid occurrences. The impact of this problem heavily depends on the studied phenomenon: *-able* tends to give English words, while *-este* and *-esque* are commonly used in other romance languages (including Latin, whose presence on the Web has to be accounted for). However, resulting documents written in these languages can be easily detected and discarded.

But the most annoying problem when dealing with Web data comes from the complete lack of information regarding the author and the context (see the discussion about Lüdeling, Evert and Baroni, 2007 in 2.1. above). This means that some detected forms are wrong for a number of reasons, and cannot be taken into account. Examples of encountered error types are:

- direct transfers from another language (most of the times the author's mother tongue). These include the results of machine translation;
- words coined for stylistic reasons (rhyme, pun, etc.);
- regionalisms or archaic words;
- plainly incomprehensible contexts, either from low-quality writing or technical jargon.

2.4 Methodological issues

Most of the problems mentioned above cannot be resolved by automated means, even if Webaffix implements some heuristics to detect non-French web pages and a few kinds of suspicious contexts. As a result, every new word form harvested by such a tool has to be manually examined. Such examinations make full use of the context: the Web page itself, but sometimes other pages on the same site. As mentioned above, the main point is to sort out legitimate constructions from errors, to get their meaning and to identify their base forms.

The number of occurrences of a given word has not been taken into account: one legitimate and one interpretable context per target word is considered sufficient. Thus, frequency was a minor issue in the works presented here. However, the search for “good” contexts often leads the linguist to examine several occurrences in order to get a clear view.

The variety of the Web confronts us to many different types of discourses, some of them making the understanding difficult. In a few cases, the authors (if identifiable) have been contacted in order to get confirmation of their intent to coin a new word. In other cases, too obscure contexts, and thus

words, have been discarded. As can be seen in the examples described hereafter, the main advantage of Web data is the spontaneity found in the productions: many interesting findings occur in forums, blogs, and other types of pages where a community of speakers freely express themselves in informal contexts. This kind of data, as will be discussed below, is one of the most interesting point of using the Web as a corpus for morphology.

Perusing data automatically extracted from the Web remains a time-consuming task, but the experiments described in the following sections will demonstrate the utility of such work.

Another important point to be mentioned is the fact that the great majority of the examples retrieved and used in the works described below was consistent with native speakers' intuition and sounded 'natural'. The advantage of the Web, in this case, is that it gives access to all (or at least the majority of) the cases we can find in relationship to a particular construction.

3 Sample studies using extensive morphology

There is no doubt that the efforts for an automatic retrieval of word forms in the huge mass of the electronic texts have brought a significant progress in our knowledge of morphology. Recent studies show that this 'extensive' approach permits new and finer generalizations, as the number of collected forms grows. In particular, extensive morphology allows us to record rare facts whose existence was uncertain due to the weakness of our intuitions. Finally, this approach makes it possible to conceive experimental ways capable of renewing, at least partially, the empirical bases of our discipline.

This section presents some of the results, in the fields of morphophonology and morphosemantics, obtained by extensive morphological studies.

3.1 New generalizations

The progress brought by the extensive morphology approach has been quick in particular in the domain of morphophonology. Concerning French, the most spectacular results are certainly those regarding the phenomena of dissimilation. As an example, we describe below how Marc Plénat and his collaborators progressively uncovered the conditions under which certain rhymes fall before the suffix *-esque*. As far as semantics is concerned, the progress is slower, but some examples, such as the derivation with *-able*, show that we can expect some questions to be completely renewed. The choice of these suffixes was not completely arbitrary: *-esque* is a suffix which can be very productively attached to proper nouns. Thus, through Web searches, we expected to find a large amount of forms not recorded in traditional lexicographic sources. *-able* was chosen because it has been often considered as a ‘simple’ element to describe, with relatively straightforward combinatorial properties and meaning, and all deviant cases had been treated as ‘marked’ or ‘peripheral’.

3.1.1 Mid vowels before *-esque*

The data which provided the largest amount of new observations are certainly those collected in the database of derived words in *-esque* developed at the ERSS research unit. This base currently includes some 3,000 different forms, each accompanied by one or more referenced examples. By comparison, the

Robert électronique and the *TLFi* contain each less than 100 words derived with this suffix. This database has constantly grown through different sources during 15 years of work: starting with a simple selective reading of books and newspapers, it then took advantage from the availability of digital corpora, and more recently from the Web search techniques previously mentioned.

This database sheds new light on the morphophonology of French in many ways. To illustrate our point, we will only relate here the appearance and progressive clarification of a problem so far ignored: the behaviour of words ending in a mid front vowel (/e, ε, ø, œ/) + a fixed consonant before *-esque*.

If we trust lexicographic sources, bases ending with such a sequence do not pose any special problem, since the few words recorded in the above mentioned dictionaries are formed by simply concatenating the suffix to the base lexeme as it appears in its free form (cf. 1):

- (1) Babel → babélesque
- Molière → moliéresque
- Raphaël → raphaélesque

The idea that identity (or similarity) between the vowel in the suffix and the final vowel of the base lexeme could cause the fall of the rhyme first appeared in the mid-90's, when the base contained some 800 items, and was put forth in Plénat (1997: 168). The new list showed in fact that a rhyme in /ε/ + a fixed consonant may sometimes disappear when the base is at least four syllables long, be it a simple (2a) or a complex (e.g. a derivative in *-eur*, 2b) word:

- (2) a. Nibelungen → nibelungesque
- Pantagruel → pantagruésque

- b. consommateur → consommatesque
 ‘consumer’
 déprédateur → déprédatesque
 ‘depredator’

The data also included one case in which the ending *-eur* disappeared in a base of only three syllables (*tirailleur* ‘sharpshooter’ → *tiraillesque*). The trisyllable *Cervantes* was also shortened in *cervantesque*, but this was not different from the most current behaviour of trisyllables ending in /s/ (cf. 3):

- (3) clitoris → clitoresque
 cosinus → cosinesque

Some years later, the retrieval of some 400 new forms did not lead to a significant advance: Plénat(2000: 32) observes the relative weakness of endings in a mid front vowel + a fixed consonant before *-esque* for long bases, but is incapable of precisely determining the factors that cause the shortening.

The current database allows a much more precise description of data: 1- it is confirmed that a rhyme in a mid front vowel + a fixed consonant may fall when the base lexeme is at least four syllables long. It is also confirmed that with shorter bases such a rhyme is maintained. (4) shows some pairs we find in the database:

- (4) Polichinelle → polichinellesque, polichinesque
 Harry Potter → harrypotteresque, harrypottesque
 Vétérinaire → vétérinaresque, vétérinesque
 ‘veterinary surgeon’
 Ordinateur → ordinateuresque, ordinalesque
 ‘computer’

However,

2- they may fall not only when the last consonant is identical to one of the consonants in the suffix (as in *cervantesque*, see above, or in *BTesque* formed on *BTS* /betɛs/ ‘technical-scientific high-school diploma’), but also if the last consonant is already represented at least once in the base (cf. 5). In other words, the *tiraillesque* case is now explained: it is the presence of two /r/ in the base which causes *-eur* to fall:

(5) ³			repeated consonant
Ben Laden	→	benladesque	/n/
colonel	→	colonesque	/l/
Internet	→	internesque	/t/
Warhammer	→	warhammesque	/r/

3- finally, they also fall even if the base only has two syllables when its last fixed consonant is identical (or almost identical) to one of the consonants of the suffix. (6) displays examples containing a sibilant or a /k/:

(6)	(Fabien) Barthez	→	barthesque
	(Louis de) Funès	→	funesque
	Cherek	→	cheresque
	name of an imaginary island		
	Other rhymes in a sibilant may disappear at the end of a disyllabic		

base (cf. 7), but less systematically than when the last vowel of a base is /ã/.

Moreover, rhymes in /k/ only fall in words with three or more syllables (cf.

8).

(7)	Phidias	→	phidiesque
	pouffiase	→	pouffiesque
	‘bitch’		
(8)	Goldorak	→	goldoresque
	‘Grendizer’, a Japanese cartoon character		
	Moby Dick	→	mobydesque

As we can see, the increase in the amount of available data is

³ *Benladésesque*, *internettesque* and *warhammeresque* are also attested.

comparable to the introduction of the microscope for natural sciences. In those areas of morphology where the observation of the few recorded forms did not show anything interesting, an enlargement of 30× revealed an appreciable amount of new facts, leading to new conclusions.

The observations above fit in a broader picture where conservative forces tend to preserve the integrity of the base lexeme and of the affix, and which in the most basic cases impose the bare concatenation of the two. However, in French, two sorts of constraints oppose to these forces: size constraints, which penalize forms of more than three syllables (Plénat, to appear), and dissimilation constraints, such as the well-known Obligatory Contour Principle (OCP), put forth, among others, by McCarthy (1986). To resume, constraints of the second type penalize the repetition in the same form of two identical or similar segments (for French cf. Plénat, 1996, 2000, Lignon and Plénat, to appear). None of these constraints can stand alone against the conservative forces, but when they operate jointly, they can trigger the truncation of the base (cf. Burzio, 2002) for a similar view of ‘alliance’ between lower ranked constraints.

3.1.2 Semantic plasticity of the derived adjectives in *-able*

Increasing the amount of available data is also a decisive factor for the semantic and categorial dimensions of morphological descriptions. The study of the *-able* derivation in French by Hathout, Plénat and Tanguy (2003) gives a good indication of the kind of advances brought by the accumulation of new data.

The adjectives derived with *-able* have often been regarded as de-

verbal adjectives with a passive meaning. In other words, their governing noun is analysed as corresponding to the direct object of the base verb or to the patient, depending on whether the relation is viewed as syntactic or thematic. This analysis has been questioned in previous studies on the *-able* derivation by Leeman and Meleuc (1990), Leeman (1992) and Anscombe and Leeman (1994), among others. These authors have put forward semantic arguments (the analysis does not capture the categorisation dimension of the derived adjectives in *-able*) as well as distributional ones (not all the verbs usable which can be passivized have a derivative in *-able*). The three studies mentioned were mainly founded on lexicographical lists and massive judgements of grammaticality. The size of the corpus was approximately 1,400 words, which roughly corresponds to the number of adjectives ending in *-able* that are listed in large dictionaries like the *Grand Larousse de la langue française*, the *Grand Robert de la langue française* or the *Trésor de la langue française (T.L.F.)*.

For their study, Hathout, Plénat and Tanguy (2003) gathered a much larger list of about 5,000 adjectives. They made use of the two collecting methods provided by Webaffix and systematically analysed the usage of some of these derived words on the Web. The study proposes a new analysis of the *-able* derivation which generalizes the traditional one while taking into account the counter-arguments mentioned above.

The collected data show that most of the *-able* derivatives have indeed a passive meaning. However, their governing noun can also represent a variety of other participants to the process. This plasticity can be easily illustrated by looking at the possible governing nouns for a derived adjective

such as *pêchable* ‘fishable’ (which, by the way, does not appear in the *T.L.F.*). Obviously, first among the things that can be said to be *pêchable* are fish and other kinds of seafood. However, places can be qualified as *pêchable* as well: (i) bodies of water (rivers, ponds, streams, etc.) and (ii) fishing spots like riverbanks, bridges, dams, etc. Depending on whether the fishing season is opened or not, whether the weather is nice or bad, seasons, days and atmospheric conditions can also be said to be *pêchable* or *impêchable* (‘unfishable’). (9) presents some examples obtained from Google.⁴

- (9) 31 Aout Eau très haute (9,7m3/s) et froide (9°C), premier **jour pêchable** depuis le 15 Aout. Quelques gobages, surtout des petits poissons, ...
sosdessoubre.free.fr/saison%202004.html
C'est vrai, la carte de pêche complète à 75€, rapportée aux nombres de **jours pêchables**, et même si ça augmente chaque année, ce n'est pas hors de prix. ...
www.achigan.net/msgforum.php?id_sujet=936&page=3
Jusqu' à 14 ça va, au delà je sors pas car le **vent** devient trop gênant voir **impêchable**. maintenant peut-être que tu as des bras trop frêles pour résister à ...
peche-en-mer-aquitai.1fr1.net/DIVERS-f5/LIENS-f18/METEO-COEF-f21/previsions-meteo-etat-de-l-ocean423.htm
Si le vent monte trop et que les **conditions** ne deviennent plus **pêchables**, plusieurs solutions s’offrent à vous :. - tout plier et attendre une accalmie ...
pechemed.free.fr/loupsurf1.htm

The authors have also found contexts where *impêchable* is governed

⁴ We do not give glosses for these examples, which only illustrate some less known uses of *pêchable*.

by fishing tackle (flies or nylon fishing lines, for instance, see (10)). Finally, not only the participants to the process, but their properties too can be characterised as *pêchable* or not: they came across examples where fish size is said to be *pêchable* (11):

(10) je remarque après quelques lancers (je peche generalement a 40 metres en etang) que mon **nylon** se met a vriller et devient **impechable**. ...

www.pechemaniac.com/forums/viewtopic_208.htm

(11) pêchable, l'ouverture du gisement à la pêche semble incompatible avec sa gestion durable. Compte tenu de la raréfaction des coques de **taille pêchable** la ...

www.reservebaiedesaintbrieuc.com/DOC/coques2002.pdf

Actually, the fishers seem to be the only participants that cannot be said to be *pêchable*.

Discovering the existence and even the proliferation of these uses which had not been identified before, helps us understand some better known adjunctive uses:

- verbs of building (12) or movement (13) yield adjectives whose governing nouns designate locations:

(12)	construire 'build'	→	constructible	un terrain constructible 'a building plot'
	bâtir 'build'	→	bâtissable	un terrain bâtissable 'a building plot'
(13)	skier 'ski'	→	skiable	une piste skiable 'a skiable run'
	rouler 'drive'	→	roulable	une piste roulable 'a drivable track'

- verbs of work yield derivative whose governing nouns designate periods of time:

(14)	ouvrer 'work'	→	ouvrable	jours ouvrables 'working days'
	travailler 'work'	→	travaillable	jours travaillables 'working days'

- verbs of condemnation yield adjectives whose governing nouns

denote charges:

(15)	pendre 'hang'	→	pendable	un tour pendable 'a rotten trick'
	enfermer 'lock up'	→	enfermable	une folie enfermable 'an insanity that would require locking up'

We also better understand that a property like the price can be said to be *abordable* 'affordable':

(16)	aborder 'approach'	→	abordable	une jupe d'un prix abordable 'a skirt at an affordable price'
------	-----------------------	---	-----------	--

3.1.3 De-nominal adjectives in *-able*

Classical descriptions of the *-able* suffixation report a number of derivatives coined from nominal bases. Gawelko (1977) had identified three small series of such adjectives, derived from names of taxes (17), vehicles (18) and titles (19):

(17)	corvée 'corvée'	→	corvéable	'liable to the corvée'
	mainmorte 'mortmain'	→	mainmortable	'mortmainable'
(18)	carrosse 'carriage'	→	carrossable	'carriageable'
	cycle 'bicycle'	→	cyclable	'bikable'
(19)	consul	→	consulable	

‘consul’		‘consulable, worthy of occupying the position of
		consul’
pape	→	papable
‘pope’		‘popeable, worthy of occupying the position of pope’

This inquiry confirms the existence of these series. Modern taxes (20), recent vehicles (21) and titles of all sorts (22) yield numerous *-able* derivatives.

(20)	TVA	→	TVable	
	‘VAT’		‘liable to VAT’	
	ISF	→	ISFable	
	‘solidarity tax on		‘liable to ISF’	
	wealth’			
(21)	jeep	→	jeepable	
	‘jeep car’		‘jeepable’	
	planche à roulette	→	planchable	
	‘skateboard’		‘skateboardable’	
(22)	recteur	→	rectorable	
	‘chief education		‘worthy of occupying the position of chief education	
	officer’		officer’	
	chaire	→	chairable	
	‘Chair’		‘worthy of being appointed to a Chair’	
	danseuse étoile	→	étoilable	
	‘prima ballerina’		‘worthy of being selected as primaballerina’	

But there is more: the study brought to light other de-nominal adjectives which match some of the remarkable de-verbal series. For instance, the adjectives derived from nouns of vehicles as in (18) and (21) can be connected to the ones derived from verbs of movement like in (11). In the same way, the de-nominal examples in (23) parallel the adjectives derived from verbs of building like the ones in (12).

(23)	piscine	→	piscinable	un terrain piscinable
	‘swimming			‘a piece of land large enough to

pool'			accomodate a swimming pool'
box	→	boxable	un garage boxable
'lock-up			'a parking space that can be
garage'			transformed into a lock-up garage'

We also can draw a parallel between adjectives derived from verbs of condemnation as in (15) and denominals like (24). The examples in (23) and (24) illustrate the fact that frequency do not play an important role in the decision to accept an example or to reject it: *boxable* occurs thousands of times and *peinable de mort* only once. Nevertheless, both are perfectly acceptable.

(24)	peine de mort	→	peinable de mort	un crime peinable de mort
	'death penalty'			'a crime that carries the death
				penalty'

Other denominals are remarkable not because of their governing nouns but because of the thematic role of their bases which can be a location as in (25) or a final state as in (26).

(25)	musée	→	muséable	une statue muséable
	'museum'			'a statue worthy of being exposed
	Matignon	→	matignonnable	in a museum'
	'French Prime			un ministre matignonnable
	Minister's			'a minister worthy of being
	official			appointed Prime Minister'
(26)	résidence'			
	frite	→	fritable	des pommes de terre fritables
	'chips'			'potatoes suitable for chips'
	fromage	→	fromageable	un lait fromageable
	'cheese'			'milk suitable for making cheese'

In summary, the inquiry clearly confirms that categorial constraints on

the *-able* derivation have a semantic origin: *-able* derivatives usually select verbs as bases because they denote processes, but when a process does not have a specific corresponding verb, a nominal base will do quite nicely.

Building up a database like the one we just mentioned is time-consuming. Even if the harvesting process has become very fast, the validation of the collected data is a lengthy philological work. But we have demonstrated that the gamble is paying off: new generalisations have been revealed in phonology, categorial constraints and semantic interpretation.

3.2 Rare facts

Looking systematically for occurrences of new words does not conflict with intuition. On the contrary, our experience suggests that speakers have very strong judgements, even for extremely rare configurations. For this kind of configuration, a Web search can confer the status of verifiable facts to intuitions that would otherwise have remained mere conjectures. A spectacular example of this type concerns the substitution of the suffix *-esque* with the ending *-este*.

Pichon (1940) had proposed an analysis of an isolated example by Verlaine (27) where he suggested that the substitution resulted from a dissimilation phenomenon that takes place after velar consonants.

(27) Silvio Pellico → sylviopelliqueste

In the next half century this conjecture has been cited many times but no other example in which *-este* was substituted to *-esque* after /k/ or /g/ had been found. Recently, Plénat *et al.* (2002) discovered half a dozen new examples, through Web searches carried out with Webaffix. For now, their

database includes about thirty new words ending in *-guese* and *-queste*, some of them being (very) well-attested, such as those in (28):

(28)	Titanic	→	titaniqueste	
	Jack Lang	→	(jack)langeste	équation Jack-langeste 'equation in the manner of Jack Lang'
	blog	→	blogeste	pause blogeste 'blogging pause'

A point of detail in the morphology of adverbs ending with *-ment* provides us with another illustration of the Web's possibilities to confirm uncertain intuitions (Plénat and Boyé, to appear). It is well-known that these adverbs are based on the stem which is also used to build the feminine of their corresponding adjective (29). The nature of manner adverbs in *-ment* as derived, rather than inflected forms, has been recently questioned (cf. Dal, 2007). Whatever their real status, however, it does not substantially change the following discussion.

(29)	<u>masculine adjective</u>	<u>feminine adjective</u>	<u>derived adverb</u>
	frais	fraîche	fraîchement
	'fresh'		
	nouveau	nouvelle	nouvellement
	'new'		
	rageur	rageuse	rageusement
	'angry'		

It is also well known that, exceptionally, adjectives ending in *-ant* and *-ent* yield adverbs in *-amment* and *-emment* respectively (30). This characteristic can be explained historically: in Old French, these adjectives were epicene words (*i.e.* they had only one form for both genders).

(30)	méchant	→	méchamment
	'mean, nasty'		
	intelligent		intelligemment
	'intelligent'		

However, not all adjectives in *-ant* and *-ent* yield adverbs in *-amment* and *-emment*. Yvon (1996:164) notes the dubious acceptability of adverbs such as (31) (his suspicion is based on a remark by Molinier (1992)).

- (31) charmant → ??charmamment
 ‘charming’
 clément → ??clémentement
 ‘mild’

Actually, we have the feeling that the derivatives in (31) are quite ungrammatical and that the adverbs derived from these adjectives must instead be coined from the feminine stems, as in (32). This unacceptability could be explained by the occurrence of two /m/ and two analogous vowels in two consecutive syllables (/mamã/): the choice of the feminine stem follows then from a dissimilative constraint.

(32)	masculine adjective	feminine adjective	derived adverb
	charmant	charmante	charmamment
	clément	clémente	clémentement

For some time, this intuition was just a hypothesis due to the scarcity of arguments which supported it. The only one available was the presence of *véhétement* among the four known exceptions to the general rule (33).

- (33) dolentement lentement présentement véhémentement
 ‘mournfully’ ‘slowly’ ‘presently’ ‘vehemently’

A recent Web search provided extra arguments in support of this view. One can find about fifty good examples of *charmamment*, even if *charmamment* appears five times too, mainly in blogs. Besides, there is one additional attestation of *charmamment* in Frantext by Albert Cohen, in his novel *Mangeclous*. As for the adverb derived from *clément*, the Web search

provided two good examples of *clémentement* but only one of *clémement* in a French-English dictionary which blindly applies the rule of grammar handbooks. Good attestations of two other exceptions (34) were discovered. To be complete, they came along two occurrences of (35), but both appear in bad (or machine) translations.

- | | | |
|------|--------------|-------------|
| (34) | aimantement | démentement |
| | ‘lovingly’ | ‘insanely’ |
| (35) | alarmamment | |
| | ‘alarmingly’ | |

Of course, these few remarks do not exhaust the subject, but these examples lead us to believe that the hypothesis was correct and that French tends to resort to adverbs in *-mantement* and *-mentement* instead of *-mamment* and *-memment*.

The previous discussion also shows that native speakers’ intuition remains essential, and that the heterogeneity of the corpora we can extract from the Web prevents us from accepting all the material without any control.

4 Conclusion

The different experiments described in this article all used an unprecedented amount of data, which was gathered through automated means, mostly on the Web. The advances achieved by these data-intensive studies on several levels have been shown. The availability of these data opens the way towards further studies: finer descriptions using more homogeneous corpora; quantitative studies; large scale comparative studies (Web data are available for a large number of languages). For now, only a few suffixes have been investigated in this new way and almost all French affixes remain to be

studied.

From a methodological point of view, Web data are of course different from what can be extracted from traditional corpora (text databases or newspaper archives), as the new word forms found on the Web can frequently be characterised as being of a lower register or from specialised terminology. The Web actually covers a variety of language usages, many of which have not been previously taken into account in wide spectrum linguistics studies (except for some airport novels or popular magazines). The informal context of the Web, especially the lack of editorial filtering, indeed gives access to more spontaneous word coinage.

In no case should this “new” kind of data be taken as a weakness for the theories deduced from it. Most spontaneous words only add to more acceptable ones in the same paradigm, and they can easily be used as a source for a new hypothesis.

Using Web material leads to new skills in detecting the nature and status of some Web pages, which is crucial in manual filtering. Generic linguistic competence is also useful in judging the level of proficiency of a given Web page's author, not to mention hunting down automated translations.

Extensive morphology certainly has good prospects. However, the rapidity in the development of this field of research will depend on the emergence of a community of extensive morphologists able and willing to share their databases. New standards have to be proposed for these databases in order to ease their reuse and their merging.

Acknowledgments

We would like to thank Marc Plénat for opening the way to extensive morphology, and for enthusiastically involving us in his works.

References

- Anscombre, J.-C. and Leeman, D. (1994). La dérivation des adjectifs en *-ble* : morphologie ou sémantique ?. *Langue française*, 103 : 32-44.
- Baayen, R. H. (1991). Quantitative aspects of morphological productivity. In G. E. Booij, J. van Marle (eds), *Yearbook of Morphology 1991*, Dordrecht: Kluwer Academic Publishers, pp. 109–149.
- Burzio, L. (2002) Surface-to-surface morphology : when your representations turn into constraints. In: P. Boucher (ed.), *Many Morphologies*. Somerville, MA : Cascadilla Press, pp. 142-177.
- Dal, G. (2007). Les adverbes de manière en *-ment* du français : dérivation ou flexion ?. In: N. Hathout and F. Montermini (eds), *Morphologie à Toulouse*. Munich: Lincom, pp. 121-149.
- Fradin, B. (1997). Esquisse d'une sémantique de la préfixation en *anti-*. *Recherches linguistiques de Vincennes*, 26: 87-112.
- Fradin, B., Dal, G., Grabar, N., Lignon, S., Namer, F., Tribout, D. and Zweigenbaum P. (to appear). Remarques sur l'usage des corpus en morphologie. *Langage*.
- Gawelko, M. (1977). *Evolution des suffixes adjectivaux en français*. Wrocław, Poland : Polska Akademia Nauk Komitet Neofilologiczny.
- Hathout, N. and Tanguy, L. (2002). Webaffix : a tool for finding and

- validating morphological links on the WWW. In: M. G. Rodríguez and C. P. S. Araujo (eds), *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain: ELRA, pp. 1799-1804.
- Hathout, N., Plénat, M. and Tanguy, L. (2003). Enquête sur les dérivés en *-able*. *Cahiers de Grammaire*, 28: 49-90.
- Hathout, N., Namer, F., Plénat, M. and Tanguy, L. (to appear). La collecte et l'utilisation des données en morphologie. In B. Fradin, F. Kerleroux and M. Plénat (eds), *Aperçus de morphologie du français*. Saint-Denis: Presses universitaires de Vincennes.
- Leeman, D. (1992). Deux classes d'adjectifs en *-ble*. *Langue française*, 96: 44-64.
- Leeman, D. and Meleuc, S. (1990). Verbes en tables et adjectifs en *-able*. *Langue française*, 87: 30-51.
- Lignon, S. and Plénat, M. (to appear). Echangisme suffixal et contraintes phonologiques. (Cas des dérivés en *-ien* et en *-icien*). In B. Fradin, F. Kerleroux and M. Plénat (eds), *Aperçus de morphologie du français*. Saint-Denis: Presses universitaires de Vincennes.
- Lüdeling, A., Evert, S. and Baroni, M. (2007). Using Web data for linguistic purposes. In: M. Hundt, N. Nesselhauf and C. Biewer (eds), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, pp. 7-24.
- Molinier, C. (1992). Sur la productivité adverbiale des adjectifs. *Langue française*, 96: 65-73.

- Namer, F. (2003). WaliM : valider les unités morphologiques par le Web. In: B. Fradin, G. Dal, F. Kerleroux, N. Hathout, M. Plénat and M. Roché (eds), *Les unités morphologiques*. Lille: Forum de morphologie, pp. 142-150.
- Pichon, E. (1940). Attache d'un suffixe à un complexe. *Le français moderne*, 8: 27-23.
- Plénat, M. (1996). De l'interaction des contraintes: une étude de cas. In: J. Durand, B. Laks (eds), *Current Trends in Phonology: Models and Methods*. Salford: ESRI, pp. 585-615.
- Plénat, M. (1997). Analyse morpho-phonologique d'un corpus d'adjectifs en *-esque*. *Journal of French Language Studies*, 7 : 163–179.
- Plénat, M. (2000). Quelques thèmes de recherche actuels en morphophonologie française. *Cahiers de lexicologie*, 77 : 27–62.
- Plénat, M. (to appear). Les contraintes de taille. In: B. Fradin, F. Kerleroux and M. Plénat (eds), *Aperçus de morphologie du français*. Saint-Denis: Presses universitaires de Vincennes.
- Plénat, M. and Boyé, G. (to appear). Le Choix des thèmes dans les dérivés désadjectivaux en français. In B. Tranel (ed), *Understanding Allomorphy. Perspectives from Optimality Theory*. London: Equinox Publishing.
- Plénat, M., Lignon, S., Serna, N. and Tanguy, L. (2002). La conjecture de Pichon. *Corpus et recherches linguistiques*, 1: 105-150.
- Resnik, P. and Elkiss, A. (2005). The linguist's search engine: an overview. In: K. Knight, H.T. Ng and K. Oflazer (eds), *Proceedings of the 43rd*

Annual Meeting of the Association for Computational Linguistics. Ann Arbor, MI: University of Michigan. pp. 33-36.

Santini, M. (2006). Identifying genres of Web pages. In: P., Mertens, C., Fairon, A., Dister and P., Watrin (eds), *Verbum ex machina. Actes de la 13e conférence sur le traitement automatique du langage (TALN 2006)*. Louvain-la-Neuve, Presses universitaires de Louvain, pp. 307-316.

Yvon, F. (1996). *Prononcer par analogie : motivation, formalisation et évaluation*. Unpublished PhD thesis, Paris: E.N.S.T.