133. Internet

Georgette Dal, Fiammetta Namer

131. Internet

Abstract

The possibility of using the Internet for morphological research constitutes a kind of Copernican revolution, which opens up new perspectives, but which has not yet been seen through to its entirety. In this article, we present a selected review of the tools and resources bound to the Internet, and we discuss the main theoretical implications of the increasing use of this medium.

1. Introduction

This chapter focuses on the use of the Internet in morphological research and on the changes such a use can give rise to. First we show what the sentence “I found it on the Internet” means for a morphologist or, more widely, for the public in general. Then we review the different tools and resources which are available for morphological studies and we examine the kinds of skills morphologists need to use them. In the third part of this article, we give examples of morphological research that use the Internet, and discuss the epistemological changes induced into the morphological field by such a medium.

2. “I found it on the Internet”

Before determining what “I found it on the Internet” means for a morphologist, we will specify what exactly is accessible on the Internet from a technical point of view.

2.1. Internet *versus* Web

The Internet is a global non-centralized system of interconnected computer networks. This worldwide interconnection of computers is designed to facilitate the sharing of information among users. The evolution of the Internet follows that of the computer sciences and telecommunications. By the end of the eighties, 100,000 computers were connected whereas nowadays the Internet is a system connecting millions of public, private, academic, business, and government networks. Through this system a large spectrum of information can be exchanged, and various devices have been developed, the most famous of which being electronic mail, instant messaging and the World Wide Web (WWW or Web).

The World Wide Web is a public hypertext system that allows users to view pages from a site (or URL). These pages, which may contain text, images, videos, music, etc., are accessed by following a hyperlink, i.e. a link interconnecting pages, or with a browser, thanks to a search engine. Search engines retrieve information from URLs that they have previously accessed. It was the emergence of the Web at the beginning of the nineties that made the Internet popular. Since then, the two terms are wrongly confused. Actually, “to look for something on the Internet” has come to mean carrying out a task on the Web, and, more precisely, on the public

or visible Web. There is, in fact, a further distinction to be made, in order to delineate the area of interest for our purpose: the visible Web and the invisible (or deep, or hidden) Web, the latter being twenty times larger than the former. The invisible Web consists of reserved access sites (discussion groups, Intranets, social networks) and dynamic pages, which are the result of database queries. The visible or surface Web, on the other hand, contains all the pages a search engine is capable of indexing, i.e. storing in its database. The public access to the Internet, that is the access to the public Web, requires the use of a commercial search engine, the most popular of which are nowadays Google, Bing and Yahoo. The estimated size of the visible Web in 2011 is approximately 50 billion indexed pages, and half of them are accessible to Google (http://www.worldwideWebsize.com/). This part of the Web is where the Internet user, and, in particular, the morphologist, can perform his/her searches.

2.2. What can be found on the Web?

On the visible Web, two types of useful resources for morphological research can be distinguished: resources which have been conceived primarily for the Internet and resources which exist in other media.

The former comprises resources such as forums or blogs. The latter concerns dictionaries, scientific or literary publications, newspaper archives, and more generally resources which are primary print resources; the Web is, then, a huge library. But it also contains resources such as slideshows and other resources conceived primarily for computer use but not for the Internet. In this case, the Web is designed as a universal repository where diverse types of resources are made available and regularly updated: e.g. lexical and syntactic databases, such as *WordNet*, *Framenet* and the *French Treebank*.

When a morphologist says "I found it on the Internet", ideally, (s)he should add "… but I could find it in print (dictionary, novel, newspaper, thesis…) if I had access to it" or "… and it is an Internet production".

In the following, we want to focus only on specific Web resources that are textual resources primarily designed on the Web for Web users, with specific tools designed to retrieve them.

3. Tools and resources

For many morphologists, searching for words on the Web comes down to performing manual queries with a commercial search engine: each word is a new query. This trivial use of the Web requires no skills in programming. Results returned by the engine (i.e. most often, the displayed sequences containing the word searched for) are stored and sorted.

In this section, we are not concerned with manual queries, but we describe several Web-related tools and resources. Web exploring tools (section 3.1.) allow us either to automatically search the Web, or to construct homemade corpora from Web data. Web-based resources (section 3.2.) are very large datasets and (pre- and/or post-processed) text corpora which are either downloadable or can be searched online.

3.1. Tools

Web searching tools are language-independent applications developed by language engineering specialists, and can be used by morphologists with no special skills in natural language processing. They include and sometimes combine at least two functionalities: the first one consists in automatically submitting queries to a (commercial) search engine, and

displaying returned results; the second one, in crawling the Web, i.e. methodically exploring Web pages and storing their content. Two kinds of basic applications are generally used to perform these tasks: Web search engine Application Programming Interfaces (APIs) and crawlers.

3.1.1. Tools for automatically searching the Web

Web Search Engine API-based tools
These applications replace human users in performing Web searches. At least two programs using such APIs have been specifically developed to make Web search for word-formation automatic: Webaffix (Hathout and Tanguy 2002) was designed to query the Altavista engine, and WaliM (Namer 2003) was initially used to work with Yahoo (and now Bing). The user provides both systems with a list of words which have to be checked online, in order to assess his/her underlying intuitions and theoretical hypotheses. For each successful query, the program displays the global word count, and, for each indexed URL, a text sequence containing the searched word. Webaffix also makes use of regular expressions (i.e. sets of symbols used to search for occurrences of text), so that any possible word that includes a given sequence is retrieved. From the returned results, the morphologist could then construct new word-formation hypotheses and assertions. Section 4 gives some examples of word-formation research making use of these two applications.

More sophisticated Linguistic Search Engines: WebCorpLive
WebCorpLive (Renouf, Kehoe and Banergee 2007) is a new version of WebCorp, still in progress. It is a Web interface containing a suite of Web search API-based tools. Its input is a user-provided sequence (word, pattern, or phrase), and this input serves as a query to a commercial engine, selected among a list of several ones: Google, Yahoo and Bing. The output format is in Key Word in Context (KWIC) style, with the search terms centered. Like Webaffix,WebCorpLive allows full pattern matching and wildcard searches; moreover, results can be displayed by date. However, as far as morphology is concerned, this application also has several limitations and drawbacks. First, it is impossible to retrieve all the results of a query since both WebCorpLive and the searched commercial engine arbitrarily impose a maximum number of output pages (for instance, Google's new policy allows 100 free queries per day only). Second, word query lists (as used with WaliM) cannot be directly used. Advantages and drawback of (an earlier version of) WebCorp to linguistic research are discussed in Lüdeling, Evert and Baroni (2007).

The use of WebCorpLive is illustrated here by the search of Italian compounds ending with <opoli>. The <*opoli> input query is submitted to Google by the interface, and returns 34 pages. For each page, all the different word forms matching the <*opoli> pattern are displayed. This is particularly interesting when a single sentence contains several matching forms, such as in Fig.131.1, lines 151–154: the collected contexts can be used to study serial effects induced in the production of complex words.

33) http://www.bookstrailers.com/2011/07/copertinaeur-1750-prezzo-eur-1575.html
Text, Wordlist, text/html, UTF8 (Content-type), 2011-08-09 (Server header)

150: E lo hanno intitolato, non a caso, '**Sprecopoli**'. Un titolo che richiama alla mente altri
151: lla storia con lo stesso suffisso: **tangentopoli**, sanitopoli, vallettopoli, bancopoli e a
152: lo stesso suffisso: tangentopoli, **sanitopoli**, vallettopoli, bancopoli e altre 'opoli'.
153: uffisso: tangentopoli, sanitopoli, **vallettopoli**, bancopoli e altre 'opoli'. Anche in
154: ntopoli, sanitopoli, vallettopoli, **bancopoli** e altre 'opoli'. Anche in questo caso,

Fig 131.1: Sample of output to the WebCorpLive query: <*opoli>

3.1.2. Using crawlers to construct corpora from the Web

Whereas the applications above automatically send queries to search engines and display the obtained results, other Web tools, called "crawlers", are designed to construct text corpora by exploring the Web content. Crawlers are tools used to explore Websites; they start from a given root page and systematically collect each page's content. For a morphologist, these tools mean that the Web can be used as a source of texts and, thus, can be used to construct homemade corpora. Instead of accessing the Web to answer his/her query, the linguist has with this solution a stable, tailored corpus at his/her disposal to work with.

A first, freely available service is *GlossaNet* (Fairon, Macé and Naets 2008) a linguistic Web crawler based on the *Unitex* linguistic development toolbox (Paumier, Nakamura and Voyatzi 2009): morphologists looking for the emergence of linguistic changes can use this tool to search in online published data in the form of RSS feeds, i.e. Web contents automatically delivered at each update.

Another ongoing experiment is being carried out by the WebCorpLive authors, who wish to bypass commercial engines and their numerous drawbacks (restrictions in the number of automated searches and in the search pattern format allowed, access policy, etc.). They are developing the *WebCorp Linguist's Search Engine* or *LSE* (Kehoe and Gee 2007), a "fully tailored linguistic search engine". Eventually, this non-commercial search engine, enriched with linguistic tools, will supersede the currently used search engines to meet the needs in morphology (as well as in other linguistic fields). The URL collection for the WebCorp LSE's future Web is carried out with a Web crawler. The final Web sample size will be approximately 10 billion words.

Other tools, such as Heritrix and BootCat, were designed to construct corpora from the Web. Both are by now freely available, each of them involved in a major Web-based corpus development project. *Heritrix* (Sigurdsson 2005) is a Web crawler and a Website archiver originally designed to realize the Internet Archive project. The *BootCaT toolkit* (Baroni and Bernardini 2004), following considerations conducted in Cavaglia and Kilgarriff (2001), is a Web crawler using, as input, a user-defined set of "seed words" (that is, terms that are expected to be typical of the user's domain of interest) to obtain thematically consistent pages from the Web and thus construct a specialized corpus. This tool is particularly interesting for the study of diastratic and diatopic variations in morphology. For instance, a lexicon in a given dialect can be used as initial seed words list in order to gather a quantitatively important corpus of texts belonging to the same diatopic variety.

It is worth noting that, unlike Web Search Engine API-based tools where no special ability in computer science is needed to use the results returned from the queries, corpora constructed with a crawler-based tool require natural language processing skills for their content to be efficiently processed. In the same way, there are two kinds of available Web-based corpora presented in section 3.2.: some are stored in online databases, and searches are made easy through user-friendly interfaces; others are downloadable, and require natural language processing skills for post-processing and exploration tasks.

3.2. Web-based corpora

Besides homemade corpora, the morphologist has at his/her disposal Web-based, ready-made, sometimes linguistically annotated, corpora. There are two kinds of sources and formats: formatted word sequences from databases (section 3.2.1.) and text corpora (section 3.2.2.). Both can be downloaded or searched online.

3.2.1. Datasets based on Google index

Two huge datasets are based on the Google search engine index. Both are stored as n-grams, i.e. sequences of 'n' consecutive words. The length of the n-grams ranges from one to five words. Single words are kept only when their Google frequency is higher than 40. This limitation is regrettable in word-formation: words attested online with lower frequencies often are the most interesting ones, for instance for those who study neology, availability of word-formation patterns, etc.

Google Index 1T 5-gram
Google Web 1T 5-gram corpus is a one-terabyte n-gram data set, often called the "1t corpus" (Brants and Franz 2006). This dataset contains word n-grams and their observed frequency counts for various European languages. Google has gathered it in 2006 for the English subset and 2008 for Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish and Swedish. For English, the counts have been generated from URL-based texts containing over 1 trillion words. For the other ten languages, the counts have been generated from texts with approximately 100 billion words for each language.

Corpora are distributed by LDC: they can be searched by the freely downloadable software Web1T5-Easy (Evert, 2010). Otherwise, for the English sub-corpus, it is possible for the morphologist to access the index set through an online interface, allowing complex queries. Several search terms enable the user to parameterize queries. For instance, entering '%gate scandal' as the search input provides a list of *-gate* words in use in 2006 and related to political scandals: e.g. *Lewinskygate*, *Kofigate*, *nazigate*, *betsygate*, *funeralgate*, or *fajitagate*.

Google books Ngram Viewer
In this second initiative based on the Google index (Michel, Shen, Aiden, Veres and Gray 2011), a corpus of digitized texts containing about 4% of all books ever printed between 1800 and 2000 has been constructed. The aim is to provide quantitative measures for the investigation of literature. The corpus is divided into files, according to the source language (at the moment: American and British English, Chinese, French, German, Hebrew, Russian, Spanish) and the n-gram length; file contents all have the same format: each line records, for an n-gram, the year it occurs, its overall occurrence number, the distinct page and book counts. There is no online search interface, so natural language processing abilities are required here to search the database: files have to be downloaded (the French 1-gram dataset alone weights 3,264 Gbytes, comprising a total of 3,271 236 828 words when unzipped) and morphologists are required to develop their own tools to get the information they need from them. Online Google book-based corpora are useful in morphology for productivity counts and diachronic comparisons. But book titles, authors' names, literary genres, etc., though they are very important features for book databases, are not (yet) available, and this dampens the interest morphology may have for this data.

3.2.2. Text corpora

Text corpora are corpora made of sets of text documents, contained in downloaded URLs automatically accessed by a crawler. Once completed, these generally large corpora are post-processed in the same way as corpora usually are, e.g. by means of grammatical or syntactic tagging (for a description of corpus pre- and post-processing, see e.g. Fradin, Dal, Grabar, Lignon, Namer, Tribout and Zweigenbaum 2008). Search patterns and other morphologist-friendly tools are available. On the other hand, query results do not cover the whole indexed Web, but only a sample of it.

We will briefly describe two such projects. Both have been developed in the framework of the WaCky program (Baroni and Bernardini 2006). They aim to provide the linguist user with corpora of a size similar to that of the Web, and of a quality comparable to that of real corpora. Both (will soon) propose a Web interface, in order to replace Google or Yahoo, with nearly comparable non-commercial engines and Web contents.

Leeds collection of Web corpora

This collection of large, representative, 'British National Corpus-sized' corpora (around 100 to 200 million tokens each) was compiled by Sharoff (2006). In 2008 it included English, Chinese, Finnish, French, German, Italian, Japanese, Polish, Portuguese, Russian and Spanish, and it can be queried via an online interface. The method used to collect corpora for each language is the following. First, a list of about 500 words is selected from among the most frequent words of the language in question. They are then used to randomly produce a list of 5 to 6000 4-gram queries; this query list is submitted to Google via a Google Search Engine API-based tool (see section 3.1.); the top ten URLs returned for each query are downloaded, and the resulting text goes through several post-processing steps, such as text encoding conversion, headings and other frames removal, part-of-speech tagging and lemmatization.

For instance, the input query <.*opoli> returns 941 results, each of them being a textual context in which the <*opoli> matching word is aligned by means of a customizable concordance tool. The output consists of a hyperlink to the source URL and the sequence containing the searched word.

<URL> ti. Il problema delle **baraccopoli** e della miseria è s
<URL>: ca nazionale, ovvero **calciopoli** , come è stato ri-ba

Fig 131.2: Sample of output to the Sharoff's interface: <.*opoli>

WaCky corpora

The WaCky project is an informal consortium of researchers who constructed four very large freely available language specific corpora from the Web for English, German, French and Italian (Baroni, Bernardini, Ferraresi and Zanchetta 2009; Ferraresi, Zanchetta, Baroni and Bernardini 2008; Ferraresi 2007). Each corpus size is approximately 2 billion words. The method used to collect Web documents is similar to that in Sharoff (2006). The main differences are that the WaCky approach consists of a BootCat-style crawl, using seed URLs, and that near-duplicate pages are detected. Each corpus (German: deWaC, Italian: itWaC, French: frWaC, English: ukWaC) has been obtained by limiting crawls to the country domains .de, .it, .fr et .uk. Initial seed words come from two distinct sources, for each language: the language basic vocabulary, and lexical items from well-established large resources, i.e. *Süddeutsche Zeitung*, *Le Monde Diplomatique, la Repubblica* newspaper corpora for, resp. deWaC, frWaC and itWaC, and the British National Corpus for ukWaC. Each corpus is tagged with parts-of-speech and lemmatized. For morphology, accessing lemmas and not only word forms is particularly useful. Among the various options, the user can parameterize his/her queries by combining part-of-speech and regular expressions, and perform distinct searches on types (lemmas) and on tokens (forms).

The use in morphology of these WaCky corpora has already been explored for Italian: the itWaC corpus (Baroni and Ueyama 2006) has been used to extract and analyze Italian N+N compounds, in order to assess the hypothesis of the distinction between relational and attributive compounding (Baroni, Guevara and Pirrelli 2007; Baroni, Guevara and Zamparelli 2009).

In sum, there are many initiatives that allow a (almost) free access to the Web as a massive source of data. Data are organized as databases or as text corpora, in a constrained sense of the 'corpus' definition:
– some applications include online Web search, with a set of customizable options enabling the morphologist to refine his/her queries,
– others propose tools to build homemade corpora,
– yet others make Web-based corpora available, sometimes tagged with parts-of-speech and lemmatized.
According to his/her goals and skills in natural language processing, the morphologist has by now at his/her disposal several solutions to make use of the Web in his/her research.

4. Why use the Internet for morphological research and what are the theoretical implications?

After having explained how to use the Internet for morphological research and having looked at what tools and resources are available, we will now go on to see why this medium, thanks to the above mentioned tools, is increasingly being used in this area and we will evaluate the theoretical implications of this development.

The first and perhaps main advantage introduced by the use of the Web in morphological studies is the amount of data simultaneously offered and, as a corollary, the main change is the perception of morphology: the massive increase in data can substantially modify the results of prior morphological studies, and can lead to new theoretical conclusions. Put more directly, the use of the Web makes it possible to formulate hypotheses about the actual morphological system.

Lindsay and Aronoff (in press) constitutes a first example. In this study, Lindsay and Aronoff explore the Web to compare the distribution of suffix pairs in English: using basic regular expression(s), they first identified all words ending in either <ic> or <ical> (or both) in Webster's 2nd International Dictionary and stripped off these final sequences to produce 11,966 unique stems. Using the Google Search API, they then execute automated queries for each stem and suffix combination (e.g. *biolog-* + {*-ic*,*-ical*}). In order to establish what they call a "productivity measure", they quantitatively compare the <ic> and <ical> ending forms for each stem pair. This experiment leads to the following conclusions: (i) in general, most stems clearly favor one suffix over the other; (ii) while *–ic* suffixation is more productive overall, *-ical* suffixation is far more productive with stems ending in <olog>. They use a similar method to examine the pair of competing suffixes *-ize* and *-ify* and conclude that, although the former is preferred in a vast majority of cases, the latter suffix tends to be attached to monosyllabic stems. The study by Lignon (in press), also based on list projection on the Web, leads to similar results for French, looking at the suffixes *-iser* and *–ifier*.

Hathout, Montermini and Tanguy (2008) and Hathout, Namer, Plénat and Tanguy (2008) present a number of recent studies in French morphology which make extensive use of data automatically collected mostly from the Web which lead to new insights on morphological phenomena. For example, Hathout, Plénat, and Tanguy (2003) gather a list of 5,000 French adjectives in *-able* from the Web. The data confirm that most of the *-able* derivatives have a passive meaning and, thus, that the noun they modify tends to be the patient of the verb-base. However, this noun can also represent a variety of other participants in the process. The authors illustrate the semantic plasticity of *-able* suffixation with the adjective *pêchable* 'fishable', which appears on the Web but not in French dictionaries. As predicted by traditional descriptions, this adjective can occur with the noun *poisson* 'fish' or one of its (co)hyponyms. But it also occurs with nouns referring to masses of water (e.g. *rivière* 'river') and geographical areas (e.g. *secteur* 'area'), temporal nouns such as *saison* 'season', *jour*

'day' or *temps* 'weather' and so on. Without the Web, this semantic plasticity may never have been observed. Extensive data have also proved their usefulness in the field of morphophonology: e.g. see Plénat (2011) and article 54 on dissimilatory phenomena in French.

Dal and Namer (2010a) provide further evidence for the usefulness of the Web to the emergence of new hypotheses. In their study, the exploration of the Web allows the authors to identify a new model of French *-ance* derivatives referring to properties in which the sequence "ance" is directly concatenated to the base, such as *blondance* (← *blond* 'blond'), *saoulance* (← *saoul* 'drunk') or *sombrance* (← *sombre* 'dark'). According to their knowledge, this model, which remains to be confirmed, has not been identified in previous studies.

As a final example, let us quote the study by Dal and Namer (2005, 2010b) which deals with French property nouns ending with *-ité* based on toponyms such as *portugalité* ('portugueseness') ← *Portugal*, and ethnic adjectives such or *africanité* ('africanness') ← *africain* ('african'). The aim of theses studies is to explain the category variation of the base, which seems unrelated to meaning. French dictionaries contain very few such derivatives; hence, no conclusion could be drawn from the observation of dictionary lists. Machine-readable newspapers do not provide sufficient word types. So, to answer this question, the authors constituted a Web-based experiment. First a list of 145 toponyms referring to very well-known countries, regions or towns was set up. Then, each member of this list was mapped onto its morphologically and/or, if any, semantically corresponding ethnic adjective(s). For instance, *Italie* was linked to *italien*; *Hongrie*, on the other side, was related to two ethnic adjectives: *hongrois* ('Hungarian') and *magyar*. With this procedure, a list of 411 candidate bases was compiled and served to automatically generate the 411 ethnic property noun counterparts ending in *-ité*. 203 of them were present on the Web. The examination of this corpus with respect to the frequency of each derivative gave rise to the following conclusions. The formal competition between toponym and ethnic adjective is correlated to constraints on the output form. The adjective-based derivative is the default case but the toponym can be chosen to avoid final sequences with identical or similar sequences at the base-affix boundary (e.g. *yéménité* is better than *yéménitité*) or to avoid *-aisité* and *-oisité* final sequences (e.g. *portugalité* is better than *portugaisité*). In some cases, the explanation seems to be a preference for well represented final sequences in the attested French lexicon (e.g. *magyarité* is better than *hongrité* – which is better than *hongroisité*), or for quadrisyllabic outputs. The experiment also shows strong avoidance of derivatives ending in <nianité> and <mianité> (e.g. *mauritanité* is better than *mauritanianité*). Without the Web, such results would never have been reached.

Another advantage in using the Web for morphological research is the possibility for morphologists to instantly verify certain intuitions or to test theoretical hypotheses. Until recently, in the vein of generative grammar and for pragmatic reasons, morphologists have had the habit of following their own intuition, possibly completed by informant surveys, to decide whether a complex lexeme is acceptable or not. With the Web, it is now possible not only to verify whether a lexeme is used or not, but also to what extent it is used, and this tool becomes an indicator of low or high acceptability.

Rainer (2003) carried out such an experiment and applied it to Italian *-issimo* suffixation, which he had studied twenty years earlier with traditional methods. This experiment confirms the acceptability contrast he made in 1983 between such formations as *contrarissimo* or *caratteristicissimo* on the one hand, and *letterarissimo* and *tragicissimo* on the other: the Web provided no occurrence (or very few) of the former and a lot of the latter.

The study by Dal and Namer (2000) constitutes another example of the use of the Web to test theoretical hypotheses. Having observed the lack of verbs with an ending in <abiliser> in dictionaries, the authors tried to explain this fact. Their explanation was that *-able* adjectives express inherent characteristics of the referents of the nouns which they modify. However, in *-iser* deadjectival verbs, the base describes the property in which the entity finds itself after the process has taken place. It logically follows that this adjective should tend to express an exogenous property (individual level property in Carlson's terminology) and not an inherent (stage level) one. Thus, the lack of *Xabiliser* verbs in dictionaries is due to a semantic incompatibility between the *Xable* entities and the requirements of *-iser* suffixation for the adjectives that it selects (*stable* 'stable' in *stabiliser* 'to stabilize' and *coupable* 'guilty' in *culpabiliser* 'to feel guilty') are not *–able* adjective in modern French). To confirm the validity of this hypothesis, they conducted an automatic Web search using 1287 <abiliser> verbs created especially for the purposes of demonstration. Only 6 of these generated terms (approximately 0.5%) got positive results, and, among them, only 3 (*commutabiliser* 'commutabilize', *portabiliser* 'portabilize', and *variabiliser* 'variabilize') presented the structure that was rejected by the authors.

More radically, the development of the use of the Web in morphological studies confirms the transition to a usage-based morphology, first initiated by the possibility of consulting authentic corpora such as newspaper archives or scientific and literary publications in electronic form (cf. article 130 on corpora). With the Web, morphologists have access to authentic productions in authentic contexts. These ecologic conditions of production and use of lexemes can in turn give rise to new studies as we saw above with *-able* suffixation.

Another important theoretical result related to the use of the Web is the awareness among morphologists that there is variation also in morphology. An example is the work carried out in the WesConVa project (cf. Dal, Lignon, Namer and Tanguy 2004). The aim of WesConVa, developed between 2004 and 2005, was to study the competition between French deverbal suffixations (mainly *-age* and *-ment* suffixations) and to compare lists of such derivatives collected in dictionaries and productions on the Web. For this experiment, the authors automatically constructed 64,000 derivatives. Among them, 9,939 were present both in dictionaries and on the Web (5,023) or only on the Web (4,916). Each derivative was manually encoded with the following information: base-verb, Web frequency, number of analyzed pages, domain of use (if any). The main result of this search was that, in 25% of the cases, no conclusion could be offered to explain the presence on the Web of a competing *–age* (*-ment*) neologism besides an already attested *-ment* (*-age*) derivative: the contexts showed no semantic difference; the domains of use were the same. The same conclusion is true for neologisms ending in *-age* and *-ment* derived from the same verb-base. More generally, the experiment invalidated certain theoretical hypotheses based on dictionaries such as Kelling (2001) which considers the difference between the two suffixations to be a question of proto-roles.

Another example of the possibility offered by the Internet for studying morphological variation can be found in Mühleisen (2010), which uses the Web to study diastratic variations in English *-ee* nouns.

In our opinion, the full implications of the awareness of the existence of variation in morphology, due to the exploration of the Web, have yet to be measured. As a first approximation, we say here that rules have to be thought as flexible (the all-or-nothing model is obsolete), with overlapping areas (see also article 53 semantic restrictions).

5. Conclusion

Using the Web for morphological research has well-known limits (Lüdeling, Evert and Baroni 2007):
– the Web is not a corpus in the strict sense as, for example, is the British National Corpus;
– the Web is constantly evolving, so most experiments are impossible to reproduce;
– most Web pages provide no information regarding authorship;
– it is necessary to distinguish between productions: some words are coined for stylistic reasons or are fun formations; others are regionalisms or archaic words; and others are direct transfers from other languages;
– the Web is 'noisy' and so morphologists and more generally linguists need to be cautious with statistical measures;
– it also contains plainly incomprehensible contexts, either due to low-quality writing or technical jargon.

The Web is not a panacea for all morphological research and does not allow us to simply dispense with the use of more traditional methods such as introspection, dictionaries or corpora in a strict sense. Despite this, we have shown several devices in which the Web content becomes a major asset in morphological analysis despite the eventual limitations: nowadays, even suspicious morphologists can not longer ignore the possibilities offered by this medium, which constitutes a Copernican revolution in the morphological field.

6. References

Baroni, Marco and Silvia Bernardini 2004 BootCaT: Bootstrapping corpora and terms from the Web. In: Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of LREC 2004*, 1313-1316. Lisbon: ELDA.

Baroni, Marco and Silvia Bernardini (eds.) 2006 *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT.

Baroni Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta 2009 The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation* 43(3): 209-226.

Baroni, Marco, Emiliano Guevara and Vito Pirrelli 2007 NN compounds in Italian: modelling category induction and analogical extension. *Lingue e linguaggio* 1(2): 263-290.

Baroni, Marco, Emiliano Guevara and Roberto Zamparelli 2009 The dual nature of deverbal nominal constructions: Evidence from acceptability ratings and corpus analysis. *Corpus Linguistics and Linguistic Theory* 5(1): 27-60.

Baroni, Marco and Motoko Ueyama 2006 Building general- and special-purpose corpora by Web crawling. In: *Proceedings of the 13th NIJL International Symposium. Language Corpora: Their Compilation and Application*, 31-40.

Brants Thorsten and Alex Franz 2006 Web 1T 5-gram version 1. Linguistic Data Consortium, Philadelphia.

Cavaglià, Gabriela and Adam Kilgarriff 2001 *Corpora from the Web*. Information Technology Research Institute Technical Report Series (ITRI-01-06), ITRI, University of Brighton.

Dal, Georgette, Stéphanie Lignon, Fiammetta Namer and Ludovic Tanguy 2004 Toile contre dictionnaires : analyse morphologique en corpus de noms déverbaux concurrents. Communication au colloque *Les noms déverbaux*, Université Lille 3, 23-25 septembre 2004.

Dal, Georgette and Fiammetta Namer 2000 GéDériF: Automatic generation and analysis of morphologically constructed lexical resources. In: Maria Gavrilidou, George Carayannis, Stella Markantontou, Stelios Piperidis and Gregory Stainhauer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation*, 1447-1454. Athens, Greece, 31 May – 2 June 2000.

Dal, Georgette and Fiammetta Namer 2005 L'exception infirme-t-elle la notion de règle ? Ou le lexique construit et la théorie de l'optimalité. *Faits de Langues* 25: 123-130.

Dal, Georgette and Fiammetta Namer 2010a Les noms en *-ance*/*-ence* du français: Quel(s) patron(s) constructionnel(s)? In: Franck Neveu, Valelia Muni Toke, Tom Klinger, Jacques Durand, Lorenza Mondada and Sophie Prévost (eds.), *Actes en ligne du 2e Congrès Mondial de Linguistique Française*, 893-907. La Nouvelle Orléans, États-Unis, 12-15 juillet 2010.

Dal, Georgette and Fiammetta Namer 2010b French property nouns toponyms or ethnic adjective: A case of base variation. In: Wolfgang U. Dressler, Dieter Kastovsky, Hans Christian Luschützky and Franz Rainer (eds.), *Variation and Change in Morphology. Selected papers from the 13th International Morphology Meeting*, Vienna February 2008, 53-73. Amsterdam/Philadelphia: Benjamins.

Evert, Stefan 2010 Google Web 1T 5-Grams Made Easy (but not for the computer). In: Adam Kilgarriff and Dekang Lin (eds.), *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, 32-40. Los Angeles: Association for Computational Linguistics.

Fairon, Cédrick, Kévin Macé and Hubert Naets 2008 GlossaNet 2: a linguistic search engine for RSS-based corpora, In: Stefan Evert, Adam Kilgarriff and Serge Sharoff (eds.), *Proceedings of LREC 2008. Workshop WAC4*, 34-39, Marrakesh: ELRA.

Ferraresi, Adriano 2007 Building a very large corpus of English obtained by Web crawling: ukWaC. MA Thesis, University of Bologna.

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni and Silvia Bernardini 2008 Introducing and evaluating ukWaC, a very large Web-derived corpus of English. In: Stefan Evert, Adam Kilgarriff and Serge Sharoff (eds.), *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, Marrakesh, 1 June 2008.

Fradin Bernard, Georgette Dal, Natalia Grabar, Stéphanie Lignon, Fiammetta Namer, Delphine Tribout and Pierre Zweigenbaum 2008 Remarques sur l'usage des corpus en morphologie. *Langage* 171: 34-59.

Hathout, Nabil, Fabio Montermini and Ludovic Tanguy 2008 Extensive data for morphology: using the World Wide Web. *Journal of French Language Studies* 18(1): 67-85.

Hathout Nabil, Fiammetta Namer, Marc Plénat and Ludovic Tanguy 2008 La collecte et l'utilisation des données en morphologie. In: Bernard Fradin, Françoise Kerleroux and Marc Plénat (eds.), *Aperçus de morphologie du français*, 267-287. Saint-Denis: Presses Universitaires de Vincennes.

Hathout, Nabil and Ludovic Tanguy 2003 Webaffix: finding and validating morphological links on the WWW. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, 1799-1804. Las Palmas de Gran Canaria, Espagne: ELRA.

Hathout, Nabil, Marc Plénat and Ludovic Tanguy 2003 Enquête sur les dérivés en *-able*. *Les Cahiers de Grammaire* 28: 49-90.

Kelling, Carmen 2001 Agentivity and Suffix Selection. In: Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG 2001 Conference*, 147-162. Stanford: CSLI Publications.

Kehoe, Andrew and Matt Gee 2007 New corpora from the Web: making Web text more 'text-like'. In: Païvi Pahta, Irma Taavitsainen, Terttu Nevalainen and Jukka Tyrkkö (eds.), *Towards Multimedia in Corpus Studies, Studies in Variation, Contacts and Change in English* 2, electronic publication, University of Helsinki.

Lignon, Stéphanie in press *-iser* and *-ifier* suffixation in French: verify data to verize hypotheses. In: Nabil Hathout, Fabio Montermini and Jesse Tseng (eds.), *Selected Proceedings of the 7th Décembrettes: Morphology in Toulouse*, München: Lincom Europa.

Lindsay, Mark and Mark Aronoff in press Natural selection in self-organizing morphological systems. In: Nabil Hathout, Fabio Montermini and Jesse Tseng (eds.), *Selected Proceedings of the 7th Décembrettes: Morphology in Toulouse*, München: Lincom Europa.

Lüdeling, Anke, Stefan Evert and Marco Baroni 2007 Using Web data for linguistic purposes. In: Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 7-24. Amsterdam/New York: Rodopi.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres and Matthew K. Gray 2011 The Google Books Team. In: Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden (eds.), *Quantitative Analysis of Culture Using Millions of Digitized Books*, *Science*, vol 331, n° 6014, 176-182.

Mühleisen, Susanne 2010 *Heterogeneity in Word-formation Patterns: a Corpus-based Analysis of Suffixation with* -ee *and its Productivity in English*. Amsterdam/Philadelphia: Benjamins.

Namer, Fiammetta 2003 Walim: valider les unités morphologiques complexes par le Web. In: Bernard Fradin, Georgette Dal, Nabil Hathout, Françoise Kerleroux, Marc Plénat and Michel Roché (eds.), *Les unités morphologiques* 3, 142-150. Villeneuve d'Ascq: Université Lille 3.

Paumier, Sébastien, Takuya Nakamura and Stavroula Voyatzi 2009 UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources. In: Sylviane Granger and Magali Paquot (eds.), *eLexicography in the 21st Century: New Challenges, New Applications,*173-175, Louvain:Presses Universitaires de Louvain.

Plénat, Marc 2011 Enquête sur divers effets des contraintes dissimilatives en français. In: Michel Roché, Gilles Boyé, Nabil Hathout, Stéphanie Lignon and Marc Plénat (eds.), *Des unités morphologiques au lexique*, 145-190. Paris: Hermès.

Rainer, Franz 2003 Studying restrictions on patterns of word-formation by means of the Internet. *Italian Journal of Linguistics* 15(1), 131-140.

Renouf, Antoinette, Andrew Kehoe and Jay Banerjee 2007 WebCorp: an integrated system for Web text search. In: Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 47-67. Amsterdam: Rodopi.

Sharoff, Serge 2006 Creating general-purpose corpora using automated search engine queries. In: Marco Baroni and Silvia Bernardini (eds.), *Wacky! Working Papers on the Web as Corpus*, Bologna: GEDIT.

Sigurdsson, Kristinn 2005 Incremental crawling with Heritrix. In: *Proceedings of the 5th I nternational Web Archiving Workshop (IWAW'05)*, Vienna: Austria.

URLs for tools and resources (August 2011):

BootCat: http://bootcat.sslmit.unibo.it/
Framenet: https://framenet.icsi.berkeley.edu/fndrupal/about
French Treebank: http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
Glossanet: http://glossa.fltr.ucl.ac.be/
Google books and ngram viewer: http://ngrams.googlelabs.com/datasets
Leeds collection of Web corpora: http://corpus.leeds.ac.uk/internet.html
Linguistic Data Consortium: http://www.ldc.upenn.edu/
Unitex: http://www-igm.univ-mlv.fr/~unitex/
Wacky: http://wacky.sslmit.unibo.it/
Web1T5:http://www.cogsci.uni-osnabrueck.de/~korpora/ws/cgi-bin/Web1T5/Web1T5_freq.perl/;
http://webascorpus.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_10_Software&subpage=FILES_50_Google_N-Gram
WebCorp Live and WebCorp Linguist's Search Engine: http://www.webcorp.org.uk/
WordNet: http://wordnet.princeton.edu/

Georgette Dal, Lille (France)
Fiammetta Namer, Nancy (France)