Applications of Computational Morphology

Applications of Computational Morphology

Béatrice Daille,[1] Cécile Fabre,[2] and Pascale Sébillot[3]
[1] *IRIN*, *University of Nantes*
[2] *ERSS*, *CNRS/University of Toulouse-le-Mirail*
[3] *IRISA*

1. Introduction

Morphological information is useful for parsing, lemmatization, and in several natural language applications: text generation, machine translation, document retrieval, etc. In this paper, we shall be less concerned with what morphological processing systems are like (cf. Sproat 1992) than with the applications of computational morphology. We first present the kind of morphological information used by NLP (natural language processing) systems. That information is inflectional or derivational and may be encoded in lexical databases or retrieved dynamically through simple processing. The best known computational systems are presented, including some new methods to automatically acquire morphological information. Secondly, several technological applications using morphological information are described. In these sections, we have chosen to favor the description of some representative works in the corresponding fields, so as to illustrate how morphology is involved; we do not lay claim to exhaustiveness. This review aims at showing how truly useful morphology is for NLP systems.

2. Morphological information used by NLP systems

This section describes the kind of morphological information that NLP systems can have at their disposal. We first present available lexical databases and explain what sort of morphological knowledge they contain as well as the way they are encoded, then detail the best known techniques that perform morphological analysis: stemming and parsing. We end this section with the description of several experiments made in order to acquire from corpora morphological knowledge that can be incorporated either in a lexical database or in a morphological system.

© 2002 Béatrice Daille, Cécile Fabre, and Pascale Sébillot. *Many Morphologies*, ed. Paul Boucher and Marc Plénat, 210–234. Somerville, MA: Cascadilla Press.

2.1. Morphology and lexical databases

2.1.1. Lexical databases with inflectional information

The lexical databases DELAS and DELAC for French (Courtois and Silberztein 1989) list simple words and compounds respectively. A simple word is defined here as a string over the French alphabet delimited by two word separators, and a compound as a string that includes at least two simple words. The DELAS dictionary contains more than 90,000 lemmas with which several types of linguistic information are associated including inflectional codes. Thanks to the systematic encoding of their inflectional properties, the lemmas can be automatically inflected. The inflection of the DELAS dictionary operates according to more than 350 different paradigms, 150 of which are verbal. For instance, all the verbs that conjugate like *amuser* (*aider*, *voler*, etc.) are associated with the code V3; all the nouns that take an *e* in the feminine and an *s* in the plural are associated with the code N32, etc. Other linguistic information is available: for a verb, whether it is transitive (+t) or intransitive (+i) and its syntactic class (for example +4 for the verb *amuser* in Table 1); for a noun, its distributional class, denoted by a semantic feature: +Conc for concrete nouns, +Hum for human ones, +Anim for animate, etc. When a word is associated with more than one inflectional class, it is represented by more than one DELAS entry as shown for the noun *cousin* in Table 1: the first entry corresponds to a person and accepts the feminine form *cousine* (code N32), while the second entry corresponds to an animal (i.e., *mosquito*) in the masculine only (code N1). The DELAF French lexicon which contains 900,000 word forms is derived from the DELAS French lexicon. The linguistic information associated with the words is the same as in the DELAS dictionary, completed with inflectional information: mood, tense, person, number. In Table 1, the code C1p for the word *amuserions* indicates that the form is conjugated in the conditional first person plural. DELAS dictionaries for English and Spanish of about 60,000 simple word entries also exist.

The DELAC dictionary contains more than 100,000 French compounds (90,000 nouns, 8,000 conjunctions, 8,000 adverbs, and 15,000 *être Prep N* constructions). The DELAC dictionary encodes information about the syntactic category of the compound, its morphosyntactic structure (*NA* for noun adjective, *NPN* for noun preposition noun, etc.), its gender or number, and how to obtain the plural form: the sign – indicates that the plural form is allowed for the corresponding item of the compound, while the sign + that the corresponding item remains invariable. In Table 1, the code –+ associated with *marche antinucléaire* indicates that only *marche* is inflected in the plural form. The French DELACF dictionary contains more than 180,000 compounds, most of them nouns (Courtois and Silberztein 1990), which have been derived from the DELAC dictionary. Each entry of the DELACF dictionary is associated with its lemma, its part of speech, and the corresponding inflectional information.

Table 1. Examples of lexical entries in the DELAF, DELAS, DELACF, and DELAC dictionaries

Dictionary	Type	Example of entry
DELAF	word-forms	*amuserions*, *amuser*. V3+t+4:C1p
DELAS	lemmas	*amuser*. V3+t+4
		cousin. N32+Hum
		cousin. N1+Anim
DELACF	word-form	*pommes de terre*, *pomme de terre*. N+NDN:fp
	compounds	*tout de suite*, *tout de suite.* ADV
DELAC	lemma	*marche antinucléaire*. N+NA:fs/–+;une
	compounds	*pomme de terre*. N+NDN:fs/–+;une
		tout de suite. ADV

The MULTEXT project (Véronis and Khouri 1995) has provided lexical lists of lemmas and inflected word-forms for four languages of the European Community: French, Italian, Spanish, and German. The word-form list contains word-forms, lemmas, and a linguistic description. The linguistic description encodes features which have been considered relevant for several languages and are based on EAGLES (Expert Advisory Group on Language Engineering Standards) recommendations for computational lexicons. The word-form dictionary for French lists 300,000 forms, including proper nouns and compounds. Each character of the linguistic description specifies a value of an attribute. For a verb, there are 7 attributes: its part of speech (V), its type (m for a main verb, a for an auxiliary verb, etc.), its mood or verbal form (i for indicative, c for conditional, etc.), its tense (p for present, f for future, etc.), its person (1, 2, 3), its number (s for singular, p for plural) and its gender (m for masculine, f for feminine, n for neuter). For a noun, there are 4 attributes: its part of speech (N), its type (c for a common noun and p for a proper noun), its gender, its number, and its case (n for nominative, g for genitive, etc.). For an adjective, there are 6 attributes: its part of speech (A), its type (f for attributive, o for ordinal, etc.), its degree (p for positive, c for comparative, etc.), its gender, its number, and its case. If an attribute does not apply, the corresponding position in the linguistic description string contains a hyphen. Compounds receive the same linguistic description as simple words. Examples of word-form entries are given in Table 2.

Table 2. Examples of word-form entries in MULTEXT dictionaries

Word-form	Lemma	Lexical description
amuserions	*amuser*	Vmcp1p-
cousin	=	Ncms-
effrayante	*effrayant*	Afpfs-
Émilie	*Émilie*	Npfs-
a_posteriori	=	Afpms-

2.1.2. Lexical databases with derivational information

CELEX (Burnage 1990) is a large multilingual database that includes extensive lexicons of English, Dutch, and German. For each language, several types of lexicons are available: lemma, word-form, abbreviation, and corpus type. In the lemma lexicon, each entry represents a full inflectional paradigm. In the word-form lexicon, entries deal specifically with one inflection. The corpus type lexicon contains strings extracted from various contemporary texts and is a representative list of real-life words, distinguished on the basis of their spelling, with detailed information about their frequency. Thus, inflectional information is present in the word-form lexicon and derivational information is in the lemma lexicon. We shall now detail the morphological information contained in those two dictionaries.

For English, the inflectional features are singular, plural, positive, comparative, superlative, infinitive, participle, present tense, past tense, 1st person verb, 2nd person verb, 3rd person verb, rare form, and head-word form. When the inflectional transformation is regular, a rule is given that allows one to compute the lemma form: for *abiding*, the rule is *@–e+ing*; by removing the suffix *-ing* and adding the suffix *+e*, one obtains the lemma of *abiding*, namely *abide*. Such affix-stripping rules are similar to those used in stemmers like Porter's (see section 2.2.1).

The lemma lexicon integrates orthographic, phonetic, morphological, syntactic, and frequency information. Derivational information is provided for each lemma through its morphological structure; for example, the morphological structure of *celebration* is *((celebrate [V]), (ion)([N|V.])[N])*. This structure first indicates that *celebration* is a noun (*[N]*). The two codes which are separated by the vertical bar | have to be interpreted as follows: before the vertical bar is the word class of the word that the affix *ion* helps to form. After the vertical bar comes the word class of the stem to which the affix *ion* is added. The dot denotes the suffix itself and marks its position, in this case after the verb. The derivational information provided by CELEX suffers from some inconsistencies. For example, the lemma *acidulated* is not decomposed and the lemma *acidulous* is analyzed as *acidulate+ous*. Such problems do not appear with Porter's stemmer, which reduces these three stems to *acidul* (see section 2.2.1).

For French, a derivational database is being developed by Dal et al. (1999). This database will associate with each lexical unit both a structural description such as the ones found in CELEX and a semantic description. Due to the cost of building such a database, it is interesting to automate it as much as possible. The two techniques retained are presented in section 2.3.

2.2. Morphological systems

2.2.1. Stemming

The aim of stemming is to group words belonging to the same family in order to reduce form variability under which a given word can appear in a text. Stemming treats inflectional and derivational affixes identically. The best known stemming algorithms were developed by Lovins (1968) and Porter (1980).

The common point between the different stemmers is to proceed in two steps: (1) the de-suffixing step which consists of removing predefined endings from words, and (2) the recoding phase which adds predefined endings to the previously obtained roots. Those two phases can be performed successively, as in Lovins' stemmer, or simultaneously, as in Porter's, which we now detail.

Porter's stemmer relies on a set of transformational rules such as *-ational* → *-ate* which transforms a word such as *relational* into *relate*. Words are coded in pseudo-syllables so as to avoid applying the stemming procedure on words that are too short. For instance, the above rule can only transform a word into a stem containing one or more such pseudo-syllables. This constraint prevents the incorrect transformation of *rational* into *rate*, for instance. Porter's stemmer only reduces suffixes; prefixes or compounds are not simplified. Through this stemmer, *comprehensibility*, *comprehensible*, *comprehension*, *comprehensive*, *comprehensively*, and *comprehensiveness* are all reduced to the same stem *comprehens*. It is important to note that contrary to morphological parsing, the stems produced by Porter's stemmer are not necessarily genuine morphemes.

2.2.2. Morphological parsing

In linguistic analysis parsing, morphological analysis usually occurs before syntactic analysis. It can be performed using lexical databases such as those presented in section 2.1 or through morphological parsing. The advantage of morphological parsing over lexical databases concerns the possibility of making predictions on either the lemma or the syntactic category of the word. Though a language such as English can be morphologically analyzed with stemming methods presented in the previous section, this is not the case for highly inflectional languages such as Finnish which require more sophisticated techniques. Koskenniemi (1983) proposes a model of two-level morphology which encompasses both morphotactics, the ordered decomposition of a word into morphemes, and morphophonemics, the alternate forms of morphemes according to the phonological context. For example, the word *specifies* is analyzed as the stem *specify* and the suffix -*s*. The addition of the suffix -*s* transforms the final *y* of *specify* into *ie*; thus *specify* and *specifie* are allomorphs. The two-level morphology model has been implemented with finite-state transducers which allow the encoding of a correspondence between two letters, one belonging to the surface form and the other belonging to the lexical form. For example, the word *specifies* receives the following representation, where + is a morpheme boundary symbol:

(1) Lexical form: *s p e c i f y + s*
Surface form: *s p e c i f i e s*

The phonological properties of the two-level model are presented in Antworth 1991. The first implementation of two-level morphology led to the KIMMO system (Karttunen 1983). KIMMO has two analytical components: the rule component and the lexicon. The rule component consists of two-level rules that account for regular orthographic alternations. The lexicon lists all morphemes (stems and affixes) in their lexical form and their morphotactic constraints. KIMMO has two processing functions: the Generator and the Recognizer. The Generator accepts as input a lexical form such as *specify+s* and returns the surface form *specifies*. The Recognizer accepts as input a surface form such as *specifies* and returns a form divided into morphemes, such as *specify+s*, plus a gloss string VERB+PRESENT+3sg. Antworth 1990 came with a version of KIMMO called PC-KIMMO that could be run under a variety of systems. In 1993, PC-KIMMO was enhanced by a third analytical component, a word grammar, that provides parse trees and feature structures (Antworth 1993). PC-KIMMO is introduced in Volk 1994 and its mechanisms are fully described in Sproat 1992.

2.3. Automatic acquisition of morphological information

All the works described in this subsection aim at automatically extracting morphological rules and/or families with minimal or no prior linguistic knowledge. In order to achieve this goal they can use a thesaurus, or a corpus of the domain they examine, or even combine the use of those two kinds of data.

Grabar and Zweigenbaum (1999), for example, describe a work within the medical domain in which a thesaurus is the starting point to begin the acquisition of word variant pairs. The available data for this study are the French version of the pathology micro-glossary of international SNOMED terminology, in which a concept is described by a term and (sometimes) synonyms, and a reference word list – words from the SNOMED micro-glossary and from the international disease classification CIM-10.

A simple means to detect two potentially morphologically related words is to study their longest common prefix. If it is long enough, the words have a high probability of belonging to the same morphological paradigm. However this coarse method can lead to noisy results. The key point here is to start the acquisition of related word pairs in a context in which one can think that two words sharing at least three starting characters are likely to be really semantically linked, i.e., the series of synonyms of the SNOMED terminology. The pairs that are obtained through this first step are considered as instances of morphological rules. For example, from *symbiose, symbiotique*, the rule *se*/*tique* is inferred. Moreover, all the pairs that share the same prefix are grouped in a morphological family. A second step consists of expanding these families: the previously detected morphological rules are applied to all words that possess one of the suffixes of these rules. However, in order to control the word variant production, only words that belong to the word reference list are generated. These new word pairs are added to the morphological families. Lastly, families that share a common word are merged.

95% of the families computed by this method are considered fully correct. For example, the two following families are obtained: (*adéno*, *adénoacanthome*, *adénocarcinome*, *adénomateuse*, *adénomateux*, *adénomatose*, *adénomatode*, *adénome*, *adénose*, *adénoïde*, *adénoïdes*) and (*abdomen*, *abdominal*, *abdominale*, *abdominales*, *abdominaux*, *abdomino*). The produced morphological rules are inflectional (*l*/*ux*; *abdominal*/*abdominaux*), derivational (*en*/*inal*; *abdomen*/*abdominal*), or related to composition – i.e., to the combination of (autonomous or non-autonomous) stems building a more complex word (*e*/*ite*; *bronche*/*bronchite* or ε[1]/*blastome*; *angio*/*angioblastome*).

Xu and Croft (1998) present a work in which word variant family (or class) acquisition is initiated through the use of a corpus. It shows that the performance of standard stemmers (Porter, etc.) that do not reflect the language use in specific corpora can be improved with the help of corpus-based word variant co-occurrence statistics; those statistics allow modification of the contents of the word classes previously generated by the stemmers. Moreover, they can also be used to create a good stemmer with no prior linguistic knowledge; that is, they can modify the contents of word classes generated by a simple n-gram approach, in which words that share (at least) n (here 3) common starting characters are grouped. The key idea here is that word forms that must actually be grouped for a given corpus co-occur in the texts of this corpus. The methodology, developed and tested on newspaper and legal document corpora, includes several steps: the unique word forms in the corpus are collected and a stemmer (Porter, KSTEM, or the n-gram one) first builds initial word families. Then co-occurrence data for word pairs in the same class are collected in 100-word windows. The measure of the significance of a word form co-occurrence is a variation of the expected mutual information measure, named *em*, that evaluates the percentage of the occurrences of two elements which are co-occurrences. Finally, two algorithms

[1] ε denotes the empty string.

which use the *em* values are applied to refine the classes. In the connected component algorithm, each class is mapped to a graph, in which an edge between two words is created only if their *em* value is greater than a threshold. The connected components of this graph form the new word classes. The optimal partition algorithm aims at finding the class partition that maximizes a gain which is also related to the *em* values. For example, in the experiment described in Xu and Croft 1998, the newspaper corpus contains 76,181 unique word variants. The Porter stemmer clusters them into 39,949 classes. The connected component algorithm refines them and 64,821 classes are obtained. Finally, the optimal partition algorithm produces 73,015 classes. The word variants eliminated by corpus analysis include both derivational and inflectional variants. For example, the initial class produced by the Porter stemmer (*abusable*, *abuse*, *abused*, *abuser*, *abusers*, *abuses*, *abusing*, *abusive*, *abusively*) is reduced to (*abuse*, *abusing*, *abuses*, *abusive*, *abusers*, *abuser*, *abused*) by the connected component algorithm. The elimination of *abusively* and *abusable* is due to the fact that during the application of the connected component algorithm, the *em* values of pairs constituted by either *abusively* or *abusable* and one of the other members of the initial class are too low (below the threshold) to allow the creation of an edge; the two words therefore do not belong to the connected component formed by the other seven words.

The relevance of these new classes is evaluated within the information retrieval framework (see section 3.5) and it is proved that globally both Porter and KSTEM as modified by the two algorithms, and the knowledge-poor n-gram stemmer using them, all show significantly better performances than the previous versions.

Jacquemin (1997) combines the use of a thesaurus and a corpus in order to automatically acquire morphological links between words. He locates the words of a multi-word term of the thesaurus that appear slightly modified with the same beginning string, in the same collocation in the corpus. For example, this method detects a morphological link between *gene* and *genic* with the help of the two collocations *gene expression* and *genic expression*.

Dal et al.'s long term aim is to develop a derivational database for French (see section 2.1.2). In Dal et al. 1999, they describe a system named DéCor which uses no prior linguistic knowledge and automatically acquires morphological links between words within a lexicon (TLFnome) with the help of the findaffix Unix tool. More precisely, DéCor builds an oriented and valued graph from TLFnome, whose nodes are lemmas and in which an edge between a base and a derived form is valued by the prefixation and suffixation rules that permit production of the derived form from the base. If the rules provide more than one base for a derived form, only the statistically most probable one is kept. Dal et al. (1999) also present DériF, a derivational parser which produces the parse tree of a tagged word with the help of a set of linguistically motivated rules,[2] an exception list, and a reference set of tagged words (TLFnome). The rules split the word into its base and affixes, and its derivational family is gathered during this step. For example, *inexplicable* gets the following analysis:

(2) [in [[expliquer $_{\text{VERBE}}$] able $_{\text{ADJ}}$] $_{\text{ADJ}}$]
(inexplicable, explicable, expliquer)

DériF also offers semantic glosses for the derived word forms that it parses. As for DéCor, results for DériF evaluated for the two suffixes *-able* and *-ité* are good (86% of correct bases found for a 2043-word test set).

To conclude this subsection, mention must be made of a slightly different work – Theron and Cloete 1997 – that presents a method to automatically acquire two-level morphology rules for morphological analyzers/generators (see section 2.2.2) from a set of pairs consisting of a word (source) and one of its inflected forms (target). Theron and Cloete's acquisition process has two phases: the segmentation of the target into morphemes, and the determination of the optimal two-level rule set with minimal discerning contexts. For the first step, string edit sequences are computed for each source–target pair – i.e., the (valued) sequences of elementary operations (deletion, insertion, etc.) which transform the source strings into the target strings. In addition to the cost of a sequence, heuristics are used in order to choose the sequence which has the highest probability of leading to a good segmentation. Then, for a given set of source–target pairs, an acyclic finite state automaton which exactly accepts the corresponding edit sequences is built. It is viewed as a directed acyclic graph whose edges are labeled with the elementary edit operations, and the determination of the segmentation for a derived form is based on the number of different edit sequences that pass through the different edges. This allows, for example, production of the following morphotactic description:

(3) target word = prefix + source + suffix
unhappier = un + happy + er

The second step first aims at defining the lexical-surface representation for each word pair. The right and left parts of the morphotactic descriptions are used as new source and target strings:

(4) source: un + happy + er
target: unhappier

[2] For example, the underlying linguistic theory states that a construction operator is defined with the help of lexical category constraints, phonological constraints, and semantic constraints.

The edit sequence *u:u n:n +:0 h:h a:a p:p p:p y:i +:0 e:e r:r* helps in producing the lexical-surface representation required by two-level rules:

(5) lexical: un+happy+er
surface: un0happi0er

Two-level rules, whose syntax is *CP operator LC _ RC*[3] (Sproat 1992) only have to be coded for character pairs in which the surface and the lexical character differ (special pairs; *y:i* in the example). The computation of the optimal rule for each special pair is done by comparing all the possible contiguous contexts (in edit sequences) of this pair against all the possible contexts of all the other pairs, with the help of a new kind of context representation which mixes left and right members of its context. The method has been tested with very good results on three different languages: English, Afrikaans, and Xhosa (a Bantu language).

3. Applications using morphological information

Morphological information is very often used in NLP applications; we present here some of the tasks in which this kind of knowledge is particularly relevant. The first one consists of automatically adding knowledge to a corpus; this may be an end in itself, but is often just the first step for an application using a corpus. We then present studies in which morphology is a means to infer lexical semantic properties for words. Morphological knowledge is also a clue for extracting the terms of a domain from corpora automatically; terminology acquisition is the third application that we describe here. Although terms may be found in a basic form in texts, they often present some variations; the fourth subsection clearly demonstrates that morpho-semantic knowledge is necessary to tackle the problem of term reformulation. The last point concerns the relevance of morphology in increasing the performances of document retrieval systems.

3.1. Linguistic annotation of corpora

A tag is a piece of information which is added to each word of a corpus in order to indicate its grammatical category, its morphological features, its syntactic role, or its semantic class (Garside et al. 1997). Categorial or part of speech (POS) tagging is the most common one. Placing a tag alongside each word in a corpus to indicate its part of speech is useful to disambiguate different grammatical uses of a word such as *works* which may be a plural noun or a singular verb. Lemmatization involves the use of tags to indicate the relationship

[3] CP: correspondence part; LC: left context; RC: right context.

of each word to its lemma (e.g., that *took* is the past tense form of *take*). It enables inflected forms of a word to be retrieved and counted along with its lemma. Three main kinds of systems exist: those which only handle POS tagging, those which handle both POS tagging and lemmatization, and those which only handle lemmatization using or not using the POS information about the word. We present below some linguistic annotation systems based on morphological information that perform either POS tagging or lemmatization. Those systems make use of dictionaries (INTEX), transformational morphology rules (CLAWS), or two-level morphology rules (FLEMM).

3.1.1. INTEX

INTEX (Silberztein 1993) is a linguistic development environment that allows users to build large-coverage finite state descriptions of natural languages and apply them to large texts. INTEX uses the electronic dictionaries DELAS and DELAC (cf. section 2.1.1) to tag and lemmatize word forms appearing in a corpus. The result of the consultation of several dictionaries produces ambiguities between simple words, and between simple words and compounds. Some of these word ambiguities can be removed by the use of local grammars that have to be written by the user. Indeed, INTEX includes several tools to edit, maintain, and debug these local grammars. But INTEX has other functionalities, such as the retrieval from texts of all utterances of a given word (grouping its inflected forms), a list of words (listed in a dictionary), a given category or, more generally, any syntactic pattern given in the form of a regular expression. The resulting index can be used to extract special patterns from the texts, to build concordances, or it can be analyzed with INTEX statistical tools.

3.1.2. CLAWS

CLAWS (Constituent Likelihood Automatic Word-tagging System) has been continuously developed since the early 1980s at UCREL (University of Lancaster, UK). CLAWS performs tagging and lemmatization using a morphological parser. Only a small dictionary, which lists frequent words and exceptions to morphological rules, is used. The program first looks the word up in the dictionary and if the word is not found, its ending is examined. In Garside et al. 1987, the list of the best-known suffixes consists of 720 endings from one to five characters. For example, the suffix *-ness* assigns the tag NOUN, and *-mp* assigns the tags NOUN or VERB except for *damp* and *lump* which are recorded in the dictionary. In cases where the ending of a word unifies with several suffixes, the longest matching word-ending is retained. Thus, for the word *available*, only *-able* ADJECTIVE is retained among the following possible tags: *-able* ADJECTIVE, *-ble* NOUN or VERB, *-le* NOUN. Special treatments are applied to defined suffixes such as the *-s* plural marker and *-er*. When several tags are assigned to a word, probabilistic tag disambiguation is performed. The lemmatization occurs after the POS tagging, which also makes use of a morphological analyzer. A set of rules is associated with each POS tag. Each rule is composed of two endings: the first one must match with the inflected form and is then stripped; the second one is added to the stem to generate the lemma. When the ending of the word form unifies with several suffixes, once again the longest one is retained. For example, the rules dealing with English verbs inflected in the third singular person of the present indicative are: *-ches*/*-ch*: *reaches* → *reach*; *-shes*/*-sh*: *flashes* → *flash*; *-ies*/*-y*: *studies* → *study*; *-sses*/*-ss*: *passes* → *pass*; *-xes*/*-x*: *relaxes* → *relax*; *-s*/: *reads* → *read*. This kind of program, which uses a morphological analyzer for POS tagging and lemmatization, handles unknown words, as opposed to programs using dictionaries only. The CLAWS system shows a success rate of 96–97% on written texts.

3.1.3. FLEMM

FLEMM is a lemmatization program developed for French by Namer (2000) that operates on a POS-tagged text and computes, by means of rules, the non-inflected form of a tagged word together with its inflectional information. This program is original in several ways, including conception, robustness, and portability. Contrary to other lemmatization programs, it does not use a general dictionary; it uses exception lists and encodes linguistic information through two-level morphological rules that make it possible to generate genuine morphemes and inflectional information. Unknown words are handled and tagging mistakes are corrected. It takes as input a text tagged either by Brill's (1992) POS tagger trained for French by Lecomte and Paroubek (1996), or by TreeTagger (Schmid 1994). The program operates in three main steps:

1. The first phase consists of checking the consistency of a word ending with the accepted endings for the POS tag associated with the word. If the ending and the POS tag are incompatible, FLEMM re-tags it. For example, if the word *micro-environnementale* receives the tag VERB, it is corrected as an ADJECTIVE because the ending *-ale* is only allowed for a limited set of verbs such as *affaler* or *avaler*.

2. The second phase segments the word according to its POS category and assigns inflectional information to it. For example, if a verb ends with *-èrent*, two suffixes could be deleted: {*-èrent, ent*}, as in *céd#èrent* → *céd#er* or *légifèr#ent* → *légifér#er*. The stripping of the suffix *-èrent* is the general case and brings the inflectional information of third person, plural, preterite. The stripping of the suffix *-ent* is considered an exception which is listed in a dictionary and brings the inflectional information of third person, plural, present.

3. The last phase handles allomorphic variations such as *é/è* in (*légifèr#ent*, *légifér#er*) by means of a set of rules in order to generate the correct lemma of the word.

FLEMM shows a success rate of more than 99% if the word has been correctly tagged for part of speech.

To conclude this survey of different annotating systems, let us mention Mikheev's (1997) study which presents a technique for a fully automatic acquisition of morphological rules that guess possible POS tags for unknown words. Mikheev addresses the problem of acquiring guessing rules for POS taggers and defines guessing rule schemata that capture ending-guessing rules used by CLAWS and morphological word-guessing rules. Morphological word-guessing rules describe how one word can be guessed given the knowledge of another word. For instance, a morphologically motivated guessing rule can state that a word is an adjective if adding the suffix *-ly* to it will result in a word.

3.2. Morphology as an aid for semantic lexicon building

A lot of natural language processing tasks require lexical semantic information. However semantic lexicons are not available for every application domain. Consequently, much research is now being conducted in order to automatically extract lexical semantic knowledge from corpora through statistical or machine learning techniques. (See the experience reported in Grefenstette 1994b or Habert et al. 1997 and Pichon and Sébillot 1997 for surveys of this field).

Following Harris' framework (Harris et al. 1989), such research tries to extract both syntagmatic and paradigmatic information, respectively studying the words that appear in the same window-based or syntactic contexts as a considered lexical unit (first-order word affinities (Grefenstette 1994a)) or the words that generate the same contexts as the key word (second-order word affinities). For example, Briscoe and Carroll 1997 and Faure and Nédellec 1999 try to automatically learn verbal argument structures and selectional restrictions; Agarwal 1995 and Bouaud et al. 1997 build semantic classes; Hearst 1992 and Morin 1997 focus on particular lexical relations, like hyperonymy. Finally, Pustejovsky et al. 1993, Bouillon et al. 2000, Sébillot et al. 2000, and Pichon and Sébillot 2000 are studies which aim at automatically obtaining more complete lexical semantic representations.

The common aim of all these works is to extract lexical semantic information from surface cues, that is, from signs (words, word co-occurrences, etc.) that are "physically" present in the texts.

Morphological cues are one of those possible surface cues. Some studies simply consider morphological information as a means to compute the lexical semantic properties of a derived word from the representation of its base and the properties of its affixes. For example, Seewald 1994 derives the process aspect of *derivation* from the *-ation* suffix; Fabre 1996 calculates the argument structures of deverbals from those of their verbal bases and from their suffixes, implementing the proposals of Di Sciullo and Williams (Williams 1981, Di Sciullo and Williams 1987).

However, Light (1996a, 1996b) considers morphology as a means to deduce lexical semantic properties of both base and derived word forms. He proposes a method which makes use of fixed correspondences between derivational affixes and lexical semantic information. His method is divided into 6 steps: (1) analyze affixes by hand to gain fixed correspondences between derivational affixes and lexical semantic information, (2) collect a large corpus of text, (3) tag it with POS tags, (4) morphologically analyze its words, (5) assign word senses to the base and derived forms of these analyses, and (6) use this morphological structure plus fixed correspondences to assign semantics to both the base senses and the derived form senses.

No information is really given on the possible automatization of the first task, though machine learning techniques (see Mitchell 1997 for a survey of this domain) applied to a training corpus would probably lead to interesting results.

After step 1, the cued lexical semantic information is axiomatized in an extension of standard first order logic. We do not present this formalization here (see Light 1996a: Chapter 6), but only consider the kind of information that can be obtained through derivational affixes and assigned to base and derived forms of the elements of a corpus on a few examples, directly inspired by Light 1996b.

The suffixes -A*ize*[4], -N*ize*, *-en*[5], -A*ify*, -N*ify* all cue a *change-of-state* feature for their derived form (*centralize, categorize, brighten, falsify, glorify*). For -A*ize*, *-en* and -A*ify*, the result state is "equivalent" to the base predicate[6] (*rstate-eq-base* feature); that is, the result of formalizing something is that it is formal. These suffixes also cue the following feature for their bases: if a state holds for some individual, then either an event described by the derived form predicate occurred previously or the predicate was always true for the individual (*is-dependent* feature); that is, if something is central then either it was centralized or it was always central.

The suffix *-ful* marks its base as abstract (*abstract*) and its derived form as the antonym of a form derived by *-less* if it exists (*less-antonym*). The suffix *-less* marks its derived form with the analogous feature *ful-antonym*.

Light experimented with 18 affixes on the Penn Treebank version of the Brown corpus and collected lexical information for 2535 bases and derived forms. The affixes globally cue nominal semantic class, verbal aspectual class, antonym relationships between words, etc. If we consider a global evaluation of

[4] The suffix *-ize* that applies to adjectival bases. The aim of the A is to distinguish it from the following suffix *-ize* that applies to nominal bases.

[5] The suffix *-en* only applies to adjectival bases.

[6] It is assumed that each word corresponds to a single semantic predicate.

the precision of the task, the features of a derived word have a 76% chance of being true, and 82% for a stem of a derived form. Recall, only established for the *re-* prefix, is 85%.

Another advantage of taking morphological cues into account is the ability to at least assign some properties to unknown derived forms. However Light is aware of the fact that morphological cueing is only one of the means that have to be used in order to extract lexical semantic information; other surface cues, like those used by the studies mentioned at the beginning of this section, are also necessary if a more complete lexical semantic representation is required.

3.3. Terminology acquisition

In terminology acquisition, morphology has been put aside for a long time. Daille (2000) integrates the handling of derivational variants of terms in the termer ACABIT and shows that a special kind of derivational variant involving a relational adjective gives informative indications of the terminological status of a candidate term.

ACABIT eases the task of the terminologist by proposing, for a given corpus, a list of candidate terms ranked from the most representative of the domain to the least by using a statistical score. Candidate terms which are extracted from the corpus belong to a special type of co-occurrence:

- the co-occurrence is oriented and follows the linear order of the text
- it is composed of two lexical units which do not belong to the class of functional words such as prepositions, articles, etc.
- it matches one of the morphosyntactic patterns of what will be called base terms, or one of their possible variations

The patterns for base terms are:

(6) **Noun1 Adj**: *emballage biodégradable*

Noun1 (Prep (Det)) Noun2: *ions calcium*, *protéine de poissons*, *chimioprophylaxie au rifampine*

Noun1 *à* Vinf: *viandes à griller*

Those base structures are not frozen and do accept several variations. The variations taken into account are:

1. Inflectional and internal morphosyntactic variants:
    - graphic and orthographic variants which gather together inflectional variants (*conservation de produit*, *conservations de produit*) and case differences
    - variations of the preposition: *chromatographie en colonne*, *chromatographie sur colonne*
    - optionality of preposition and article: *fixation azote*, *fixation d'azote*, *fixation de l'azote*
2. Internal modification variants: insertion inside the base term structure of a modifier such as the adjective inside the *Noun1 (Prep (Det)) Noun2* structure: *lait de brebis*, *lait cru de brebis*
3. Coordinational variants: coordination of base term structures, as in *alimentation humaine*, *alimentation animale et humaine*
4. Predicative variants: the predicative role of the adjective, as in *pectine méthylée*, *ces pectines sont méthylées*

The handling of derivational variants allows grouping of synonymous nominal phrases with *Noun Adj* and *Noun Prep Noun* structures referring to a unique concept, such as:

(7) *acidité sanguine* 'blood acidity' / *acidité du sang* 'acidity of the blood'
conquête spatiale 'space conquest' /
    conquête de l'espace 'conquest of space'
débit horaire 'hourly rate' / *débit par heure* 'rate per hour'
expérimentations animales 'animal experimentation' /
    expérimentations sur les animaux 'experimentation on animals'

These *Noun Adj* candidates involve a so-called relational adjective which has the properties of being denominal – morphologically derived from a noun by adjunction of a suffix – and paraphrasable by a prepositional phrase. The evaluation of the informative status of candidate terms involving a relational adjective, with the help of a thesaurus of the domain, shows that those candidates are 10 times more likely to be terms than their equivalents in *Noun1 Prep Noun* structure.

3.4. Detecting term variation

Detecting paraphrases is an issue for information retrieval (see section 3.5) and terminology (see section 3.3). Terminology acquisition or document retrieval applications need mechanisms to determine how information can be preserved in different linguistic realizations. Jacquemin and Tzoukermann (1999) show that morpho-syntactic variation may affect the terminology in such a way that the recognition of the occurrences of the terms in the texts is impossible, unless the system integrates some procedures to identify variants. We present an experiment (Fabre 1998, Fabre and Jacquemin 2000) which aims at automatically detecting verbal variants of nominal terms, focusing on the recognition of variants of *Noun Prep Noun* (*N P N*) terms, where one of the nouns of the initial term is morphologically related to the verb of the verbal phrase, as in:

(8) *porteur du gène* ‘gene carrier’ / *portant le gène* ‘carrying the gene’
calcul de coefficient ‘computation of the coefficient’ /
coefficient a été calculé ‘coefficient has been computed’

The following phrases, however, cannot be considered as variants:

(9) *méthode d’utilisation* ‘method of use’ / *utiliser une méthode* ‘use a method’

Such alternations are frequent in texts. Systems devoted to terminology acquisition have mainly focused on nominal terms. Relying on morphological derivation is a means to move on to the treatment of verbs.

A first step, carried out by Jacquemin (1996) by means of his system FASTR (FAst Syntactic Term Recognizer), consisted of detecting these variants on a morpho-syntactic basis. Phrases including a verb form are considered as variant candidates if the verb form is morphologically linked to one of the two nouns, and if this verbal phrase matches certain syntactic patterns. Consider, for example, two of the seven metarules that perform verbal variant recognition:

(10) **metarule 1**: N1 P2 N3 → V1 (Av ? (P ? D | P) A ?) N3
< V1 der ref > = < N1 ref >

(11) **metarule 2**: N1 P2 N3 → V3 ((P? D) | (P D ?) A?) N1
< V3 der ref > = < N3 ref >

The rule in (10) applies when, for a nominal phrase of the form *N1 P2 N3*, the parser finds in the text a phrase framed by a verb and the same noun *N3*, in which there exists a morphological link between the verb (*V1*) and the first noun (*N1*). Some parts of speech may intervene between the verb and the noun (? marks optionality, | disjunction). Examples of pairs extracted by this metarule are shown in (12):

(12) a. *stabilisation de prix* ‘price stabilization’ →
stabiliser leur prix ‘stabilize their price’
(pattern: *V1 N3*)

b. *introduction du gène* ‘introduction of the gene’ →
introduit dans le gène ‘introduced in the gene’
(pattern: *V1 P N3*)

Compared to the rule in (10), the rule in (11) allows an inversion of the words, as in (13):

(13) a. *critère d'évaluation* 'evaluation criterion' →
évalué selon les critères 'evaluated with the criteria'

b. *expérience de utilisation* 'experience of use' →
utilisait une expérience 'used an experience'

First, the results of the application of these metarules are examined on 1000 pairs extracted from a technical corpus. It appears that pairs that are extracted by means of such morpho-syntactic criteria are not necessarily variants of each other: they do not always convey similar information content. Comparing pairs of variants (as in examples (12a) and (13a)) and non-variants (as in examples (12b) and (13b)) leads to the conclusion that the frontier between good and bad variants lies between those that preserve the argument relation between the two content words and those that disrupt it. In example (12a) the object relation between the two nouns is maintained between the verb and the noun, whereas in example (12b) the verbal phrase shows a locative relation. In the second phase, metarules are enriched to reject impossible argument configurations (marked with the symbol * in what follows). Additional constraints are determined to filter out non-variants, such as (14), (15), and (16):

(14) **For metarule 1**: N1(action deverbal) *de* N3 * → V1 P N3

In a nominal phrase of the form *N1 de N3*, where *N1* is an action deverbal, the argument relation is basically an object relation. Consequently, the verb and the noun cannot be separated by a preposition (as in example (12b)).

(15) **For metarule 1**: N1(action deverbal) *de* N3 * → V1 (passive) N3

Similarly, the object relation cannot be expressed by a verbal phrase in which the verb is a passive (as in *détermination de dimension* 'determination of dimension' * → *déterminée par les dimensions* 'determined by the dimensions').

(16) **For metarule 2**: N1 prep N3 (action deverbal) * → V3 N1

When the first noun is an argument of the second one (*N3* is an action deverbal), the argument relation is basically a non-thematic relation. The *NP* cannot be paraphrased by a transitive relation in the verbal phrase (as in example (13b)).

These heuristics are based both on linguistic results on *Noun Prep Noun* in French (Bartning 1990, Fabre 1996) and on corpus observation. The refinement of the metarules implies the encoding of additional lexical features, such as transitivity/ergativity for the verbs, and deverbal properties for the nouns. This experiment shows that terminology acquisition and document retrieval systems are not simply concerned with stemming issues, and that morpho-semantic knowledge is also needed to deal with the problem of reformulation.

3.5. Document retrieval

The role of a document retrieval system is to deliver the documents that match a query that is addressed to a textual database. The best known example of such techniques is search engines on the Web. The effectiveness of the retrieval procedure is measured via two indices: precision, which shows the proportion of good solutions within the set of texts that the system has selected; and recall, which reflects the proportion of solutions within the textual database that the system delivers to the user. Basically this task is performed by statistical techniques, but adding linguistic knowledge has proven to be – to some extent – a means to improve the performance of such systems. Two directions are currently being investigated: first, linguistic analysis helps to determine which linguistic units are likely to be the best content descriptors of a text (indexing vocabulary); second, linguistic knowledge is used to cope with the problem of word variants by introducing flexibility in the matching procedure, so that different linguistic realizations conveying the same information content may be grouped and considered equivalent.

Morphological analysis is commonly used to relate equivalent terms in order to increase recall rate. Contrary to more sophisticated linguistic methods involving syntactic and semantic analysis, the morphological level is compatible with knowledge-poor, domain-independent approaches. In retrieval systems, morphological analysis is usually performed to group terms by using stemming algorithms (in particular Porter's stemmer). It allows close grouping of related words such as *computer*, *computing*, *computational* under the same index COMPUT. The influence of stemming in an information retrieval system has been evaluated by several experiments for English (Lennon et al. 1981, Harman 1991, Hull 1996, Xu and Croft 1998). There has been much debate concerning the effect of the stemming procedure on the retrieval task. While Lennon et al. (1981) and Harman (1991) conclude that stemming does not improve results – and in particular that sophisticated procedures of morphological analysis give similar results to very basic solutions – Hull (1996) and Xu and Croft (1998) come to the opposite conclusion. For example, in Xu and Croft's work, whose principles have already been presented in section 2.3, query expansion using word classes first generated by Porter, KSTEM, and an n-gram stemmer and then filtered through the use of word co-occurrences discovered in two kinds of corpora (legal documents and newspapers) leads to significantly better performances, whatever the type of the queries (word- or phrase-based). On the legal document corpus, the Porter stemmer initially builds 27,117 classes with an average size of 1.84 from the 49,964 unique word variants. The connected component algorithm breaks them into 40,215 classes (average size: 1.17), and the optimal partition algorithm produces 46,632 classes (average size: 1.07). The expansion factor of a stemmer is defined as the number of words in an expanded query set by the stemmer over the number of words in the unexpanded query set. For the Porter stemmer, its value is 5.5. For example, this stemmer expands the word *constitute* to (*constitute*, *constituted*, *constitutent*, *constitutents*, *constitutes*, *constituting*, *constitution*, *constitutional*, *constitutions*, *constitutive*, *constituttional*). The connected component algorithm leads to a 2.9 expansion factor, and the optimal partition algorithm to a 2.6 expansion factor. Less expansion not only means faster retrieval but also results in better retrieval effectiveness (an average gain of +1.3% for the first and +2.3% for the second algorithms compared to Porter). For a query containing the word *constitute*, removing the derivational variants *constitution, constitutional*, etc. significantly improves the query. Daille and Jacquemin (1998) compare two kinds of approaches for controlled indexing: one approach uses a lexical database and the other uses a stemming procedure. The evaluation shows that even if results are better with a lexical database, they are not very far from those obtained with a stemmer. Even if the results of morphological analysis on retrieval performance have not yet been firmly assessed, stemming is now widely applied in document retrieval systems for English.

In parallel, other experiments show that stemming improves efficiency more significantly for languages that have a greater degree of morphological complexity, such as French or Italian. For French, Gaussier et al. (1997) show that stemming enhances the precision by 18% while Jacquemin and Tzoukermann (1999) demonstrate that term grouping increases coverage up to 30%. Research in the document retrieval area is more and more concerned with multilingual indexing methods. Works on the morphology of French, German, Italian, Dutch, Chinese, etc. are being developed to cope with the problem of handling multilingual collections of texts. Wechsler et al. (1997) present the indexing of the Intranet of Zurich University, where 4 languages coexist: English, German, French, and Italian. For German, the authors adopt a dictionary-based approach to perform the decomposition of compounds. Next, a word reduction module is applied for the detection of words not found in the lexicon. For French and Italian, stemming algorithms are developed, adapting Porter's approach. Here is an example of the stemming results for French:

(17) Paris, 1er jan (ats/afp) Cinq cent mille personnes se sont retrouvées
samedi minuit sur les Champs-Élysées pour célébrer la nouvelle année.
Pour la deuxième année consécutive, l'avenue parisienne était interdite
aux véhicules.

Pari 1er jan at afp Cinq cent mille person se sont retrouver
samedi minuit sur le Champ Élyser pour célébrer la nouvel anner.
Pour la deux anner consécutif l avenue parisien était interdite
al véhicule.

What is shown by such an experiment is that algorithms that do not always generate linguistic units (for example, *année*/*anner*, *aux*/*al* in the French text) can nevertheless be useful for document indexing. In particular, this example illustrates the fact that document retrieval systems must address the problem of language-neutral words, especially proper nouns (*Paris*, *Champs-Élysées*), that must be filtered out to escape morphological analysis. Over-generation is the main issue in this context.

4. Conclusion

Morphological knowledge is widely used for the automatic processing of natural language: linguistic resources enriched with morphological information are currently being developed, such as corpora and lexical databases. Not only inflectional but also derivational information have proven to be helpful in several areas of NLP. With the development of multilingual NLP, morphological descriptions of various languages are needed. Even applications that require little linguistic knowledge, such as terminology acquisition and information retrieval, include some amount of morphological treatment, in order to deal with crucial issues such as the recognition of unknown words, the acquisition of term variation, or the detection of paraphrases. The part played by morphology in NLP technology can also be explained by the fact that sophisticated linguistic models are not always required, and that performances are sometimes improved – particularly as far as English is concerned – by morphological analysis based on very simple assumptions regarding word construction phenomena. But the handling of derivational morphology in NLP systems brings more than simple improvements of these systems; it also provides cues to deduce linguistic properties of words such as their lexical semantics, or to predict their informative status. In knowledge acquisition from corpora, the role of derivational morphology still has to be investigated.

References

Agarwal, Rajeev. 1995. *Semantic feature extraction from technical texts with limited human intervention*. Ph.D. dissertation, Mississippi State University.

Antworth, Evan L. 1990. PC-KIMMO: A two-level processor for morphological analysis. *Summer Institute of Linguistics Technical Report* 16.

Antworth, Evan L. 1991. Introduction to two-level phonology: Notes on linguistics. *Summer Institute of Linguistics Technical Report* 53.

Antworth, Evan L. 1993. Glossing text with the PC-KIMMO morphological parser. *Computers and the Humanities* 26: 475–484.

Bartning, Inge. 1990. Les syntagmes binominaux en *de* les types interprétatifs subjectifs et agentifs. In *Actes du Dixième Congrès des Romanistes Scandinaves*, ed. L. Lindvall, 20–34. Lund: Lund University Press.

Bouaud, Jacques, Benoît Habert, Adeline Nazarenko, and Pierre Zweigenbaum. 1997. Regroupements issus de dépendances syntaxiques en corpus: catégorisation et confrontation avec deux modélisations conceptuelles. In *Ingénierie des connaissances: évolutions récentes et nouveaux défis*, ed. Jean Charlet et al., 275–290. Paris: Eyrolles.

Bouillon, Pierrette, Cécile Fabre, Pascale Sébillot, and Laurence Jacqmin. 2000. Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL: Traitement Automatique des Langues* 41.2: 367–393.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 152–155. Morristown, NJ: Association for Computational Linguistics.

Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, 356–363. Morristown, NJ: Association for Computational Linguistics.

Burnage, Gavin. 1990. *CELEX: A guide for users*. Nijmegen: Center for Lexical Information, University of Nijmegen. http://www.kun.nl/celex/.

Courtois, Blandine and Max Silberztein. 1989. Les dictionnaires électroniques DELAS et DELAC. *Linguistica Communicatio* 1.1: 41–47, 1.2: 64–80.

Courtois, Blandine and Max Silberztein, eds. 1990. *Dictionnaires électroniques du français*. Paris: Larousse.

Daille, Béatrice. 2000. Morphological rule induction for terminology acquisition. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 215–221. San Francisco: Morgan Kaufmann Publishers.

Daille, Béatrice and Christian Jacquemin. 1998. Lexical database and information access: A fruitful association. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC '98)*, ed. A. Rubio et al., 669–673. Paris: ELRA.

Dal, Georgette, Nabil Hathout, and Fiammetta Namer. 1999. Construire un lexique dérivationnel: théorie et réalisations. In *Actes de la VIe conférence sur le Traitement Automatique des Langues Naturelles (TALN '99)*, 115–124. Paris: Talana, Université Paris 7.

Di Sciullo, Anna-Maria and Edwin Williams. 1987. *On the definition of word*. Cambridge, MA: MIT Press.

Fabre, Cécile. 1996. *Interprétation automatique des séquences binominales en anglais et en français: application à la recherche d'informations*. Ph.D. dissertation, University of Rennes 1, France.

Fabre, Cécile. 1998. Repérage de variantes dérivationnelles de termes. Technical report, Carnets de grammaire, Équipe de Recherche en Syntaxe et Sémantique UMR 5610, CNRS et Université de Toulouse-Le Mirail.

Fabre, Cécile and Christian Jacquemin. 2000. Boosting variant recognition with light semantics. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 264–270. San Francisco: Morgan Kaufmann Publishers.

Faure, David and Claire Nédellec. 1999. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Knowledge Acquisition, Modeling and Management: 11th European Workshop, EKAW '99*, ed. Dieter Fensel and Rudi Studer, 329–334. Berlin: Springer-Verlag.

Garside, Roger, Geoffrey Leech, and Anthony McEnery. 1997. *Corpus annotation: Linguistic information from computer text corpora*. London: Longman.

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson, eds. 1987. *The computational analysis of English*. London: Longman.

Gaussier, Éric, Gregory Grefenstette, and Maximilian B. Schulze. 1997. Traitement du langage naturel et recherche d'informations: quelques expériences sur le français. In *Actes des premières journées scientifiques et techniques du réseau francophone de l'ingénierie de la langue de l'AUPELF-UREF (FRANCIL '97)*, 9–14. Avignon: AUPELF-UREF.

Grabar, Natalia and Pierre Zweigenbaum. 1999. Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Actes de la VIe conférence sur le Traitement Automatique des Langues Naturelles (TALN '99)*, 175–184. Paris: Talana, Université Paris 7.

Grefenstette, Gregory. 1994a. Corpus-derived first, second and third-order word affinities. In *Proceedings of EURALEX '94*, ed. W. Martin et al., 279–290. Amsterdam: Vrije Universiteit.

Grefenstette, Gregory. 1994b. *Explorations in automatic thesaurus discovery*. Dordrecht: Kluwer.

Habert, Benoît, Adeline Nazarenko, and André Salem. 1997. *Les linguistiques de corpus*. Paris: Armand Collin/Masson.

Harman, Donna. 1991. How effective is suffixing? *JASIS: Journal of the American Society for Information Science* 42: 7–15.

Harris, Zelig, Michael Gottfried, Thomas Ryckman, Anne Daladier, Paul Mattick, Jr., Tzvee N. Harris, and Suzanna Harris. 1989. *The form of information in science: Analysis of an immunology sublanguage*. Dordrecht: Kluwer.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1992)*, 539–545. Morristown, NJ: Association for Computational Linguistics.

Hull, David A. 1996. Stemming algorithms: A case study for detailed evaluation. *JASIS: Journal of the American Society for Information Science* 47.1: 70–84.

Jacquemin, Christian. 1996. A symbolic and surgical acquisition of terms through variation. In *Connectionist, statistical and symbolic approaches to learning for natural language processing*, ed. S. Wermter, E. Riloff, and G. Scheler, 425–438. Heidelberg: Springer.

Jacquemin, Christian. 1997. Guessing morphology from terms and corpora. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ed. Peter Willett and A. Desai Narasimhalu, 156–165. New York: Association for Computing Machinery.

Jacquemin, Christian and Evelyne Tzoukermann. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In *Natural language information retrieval*, ed. Tomek Strzalkowski, 25–74. Boston: Kluwer.

Karttunen, Lauri. 1983. KIMMO: A general morphological processor. *Linguistic Forum* 22: 163–186.

Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Ph.D. dissertation, University of Helsinki.

Lecomte, Josette and Patrick Paroubek. 1996. Le catégoriseur d'Eric Brill: mise en œuvre de la version entraînée à l'INaLF. Technical report, CNRS-INALF, Nancy, France.

Lennon, Martin, David S. Pierce, Brian D. Tarry, and Peter Willet. 1981. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science* 3: 177–183.

Light, Marc N. 1996a. *Morphological cues for lexical semantics*. Ph.D. dissertation, University of Rochester.

Light, Marc N. 1996b. Morphological cues for lexical semantics. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 25–31. San Francisco: Morgan Kaufmann Publishers.

Lovins, Julie B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11: 22–31.

Mikheev, Andrei. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23.3: 405–423.

Mitchell, Tom M. 1997. *Machine learning*. San Francisco: McGraw-Hill.

Morin, Emmanuel. 1997. Extraction de liens sémantiques entre termes dans des corpus de textes techniques: application à l'hyponymie. In *Actes de la IVe conférence sur le Traitement Automatique des Langues Naturelles (TALN '97)*, 178–182. Grenoble: Communication Langagière et Interaction Personne-Système.

Namer, Fiametta. 2000. FLEMM: un analyseur flexionnel du français à base de règles. *TAL: Traitement Automatique des Langues* 41.2: 523–548.

Pichon, Ronan and Pascale Sébillot. 1997. Acquisition automatique d'informations lexicales à partir de corpus: un bilan. *INRIA Research Report* 3321.

Pichon, Ronan and Pascale Sébillot. 2000. From corpus to lexicon: From contexts to semantic features. In *PALC '99: Practical Applications in Language Corpora*, ed. Barbara Lewandowska-Tomaszczyk and Patrick James Melia. Frankfurt: Peter Lang.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14: 130–137.

Pustejovsky, James, Peter Anick, and Sabine Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics* 19.2: 331–358.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, ed. D. Jones, 44–49. Manchester: UMIST.

Sébillot, Pascale, Pierrette Bouillon, Vincent Claveau, Cécile Fabre, Laurence Jacqmin, and Jacques Nicolas. 2000. Apprentissage en corpus de couples nom-verbe pour la construction d'un lexique génératif. In *Actes des JADT 2000 (Journées d'Analyse de Données Textuelles)*, 205–212. Lausanne: École Polytechnique Fédérale de Lausanne.

Seewald, Uta. 1994. Traitement orienté-objet de la morphologie dérivationnelle. *TAL: Traitement Automatique des Langues* 35.2: 77–91.

Silberztein, Max. 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson.

Sproat, Richard. 1992. *Morphology and computation*. Cambridge, MA: MIT Press.

Theron, Pieter and Ian Cloete. 1997. Automatic acquisition of two-level morphological rules. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 103–110. Morristown, NJ: Association for Computational Linguistics.

Véronis, Jean and Liliane Khouri. 1995. Étiquetage grammatical multilingue: le projet Multext. *TAL: Traitement Automatique des Langues* 36.1–2: 233–248.

Volk, Martin. 1994. Introduction to PC-KIMMO. Presented at the 1994 North Texas Natural Language Processing Workshop. <http://www.ifi.unizh.ch/~volk/LexMorphVorl/Kimmo.Intro.html>.

Wechsler, Martin, Paraic Sheridan, and Peter Schuble. 1997. Multi-language text indexing for internet retrieval. In *Proceedings of the 5th RIAO Conference on Computer-Assisted Information Searching on the Internet*, 217–232. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaires.

Williams, Edwin. 1981. Argument structure and morphology. *The Linguistic Review* 1: 81–114.

Xu, Jinxi and Bruce W. Croft. 1998. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 16.1: 61–81.

Notes to authors:

We changed the Courtois and Silberztein 1989 reference from the conference proceedings to the journal appearance so we could cite the page numbers. (We could not find enough publication details about the conference proceedings.) Is that ok? It appears to be the same paper, and the journal will probably be easier for readers to find if they want to read the paper.

For all other conference proceedings references, we have filled in page numbers, volume editors, and publisher information (often the conference itself seems to be the official publisher) as best we can. If you see any errors or want to cite a conference presentation rather than the printed proceedings, please mark that reference appropriately.