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 of automatically acquiring 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.

 Many Morphologies, ed. Paul Boucher, 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 compounds	*pommes de terre*, *pomme de terre*. N+NDN:fp
		tout de suite, *tout de suite*. ADV
DELAC	lemma compounds	*marche antinucléaire*. N+NA:fs/–+;une
		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.

The complete article appears in *Many Morphologies*, edited by Paul Boucher, published in 2002 by Cascadilla Press. To purchase *Many Morphologies* or to find out more about the book, please visit our web site at http://www.cascadilla.com/manym.html or contact us:

Cascadilla Press
P.O. Box 440355
Somerville, MA 02144
USA

phone: 1-617-776-2370
fax: 1-617-776-2271
e-mail: sales@cascadilla.com
http://www.cascadilla.com