Erschienen in: Polibits. An open access research journal on Computer Science and Computer Engineering, Jg. 39 (2009), S. 41–48

SMM: Detailed, Structured Morphological Analysis for Spanish

Cerstin Mahlow and Michael Piotrowski

***Abstract*—We present a morphological analyzer for Spanish called *SMM*. SMM is implemented in the grammar development framework *Malaga*, which is based on the formalism of *Left-Associative Grammar*. We briefly present the Malaga framework, describe the implementation decisions for some interesting morphological phenomena of Spanish, and report on the evaluation results from the analysis of corpora. SMM was originally only designed for analyzing word forms; in this article we outline two approaches for using SMM and the facilities provided by Malaga to also generate verbal paradigms. SMM can also be embedded into applications by making use of the Malaga programming interface; we briefly discuss some application scenarios.**

***Index Terms*—Natural language processing, morphology, Malaga, Spanish.**

I. Introduction

MORPHOLOGY is one of the core processes of language. By applying the rules for inflection, derivation, and compounding, humans are able to create and understand the word forms required to communicate, including the creation of new words from existing words. To understand an utterance in some language we have to know the rules of syntax and morphology, as these are essential prerequisites for dealing with semantics or even pragmatics.

>From the point of view of computational linguistics, morphological resources form the basis for all higher-level applications. A morphological component should thus be capable of analyzing single word forms as well as whole corpora, and it should provide detailed analyses describing the relevant morphological processes. For evaluation purposes, it should also provide statistical information on speed, accuracy, etc. when analyzing large corpora.

The *Malaga* system provides a framework that supports both the development of morphological components and their application. In section II, we will give a short overview of the Malaga framework and the underlying formalism of *Left-Associative Grammar*. In the rest of this article, we will then present a specific application of Malaga, a morphological component for Spanish – the *Spanish Malaga Morphology (SMM)*.

In section III, we describe some important morphological phenomena of Spanish and present a number of principles for handling these phenomena, which guided the design of SMM. In section IV, we describe the implementation of SMM. Section V reports on the performance of SMM on two corpora. This is followed by an overview of related work (section VI) and a discussion of the use of SMM in a variety of applications. Section VIII summarizes the properties and specific advantages of SMM and outlines future work.

Manuscript received January 30, 2009. Manuscript accepted for publication March 16, 2009.

The authors are with the Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, 8050 Zurich, Switzerland; e-mail: mahlow@cl.uzh.ch; mxp@cl.uzh.ch.

II. Malaga and Left-Associative Grammar

Malaga is a software package for the development and application of morphology and syntax grammars based on the *Left-Associative Grammar* (LAG) formalism [1], providing a specialized programming language and associated development tools.

Left-Associative Grammar is based on non-deterministic finite automata. As implemented in Malaga, the analysis states are augmented by arbitrarily complex feature structures. In a morphology grammar, the symbols read from the input are allomorphs. The feature structures allow to store all available information about the involved allomorphs and the values resulting from the concatenation of these allomorphs. For the presentation of analysis results the information can be filtered to show only the features needed for a certain purpose.

Morphological components implemented in Malaga are based on the *allomorph approach*, which we will briefly describe in section IV-A.[1] Thus, the run-time lexicon used by Malaga grammars is an *allomorph lexicon* generated from a base form lexicon by applying allomorphy rules at compile time.[2]

Malaga is able to process text in UTF-8 encoding. Besides the morphological component for Spanish described in this paper, a number of Malaga grammars for morphological and syntactical analysis of English, Finnish, German, Italian, and Korean have been created, both at the University of Erlangen (Germany), where Malaga was originally developed, and elsewhere.

Malaga is freely available under the GNU Public License (GPL). For the work described in this paper we used Malaga version 7.12 on Mac OS X and Linux.[3]

[1] See [2] for a comparison of methods for morphological analyzers.

[2] See Björn Beutel: *Malaga. A Grammar Development Environment for Natural Languages*, http://home.arcor.de/bjoern-beutel/malaga/ [last access 2009-02-04].

[3] We have also used this and earlier versions of Malaga on various versions of Solaris, HP-UX, and NetBSD.

III. Spanish Morphology

Spanish, an Ibero-Romance language, is one of the most widely-spoken languages of the world. On the grounds of its rich verbal morphology it can be classified as an inflecting language; however, almost all of the noun inflections have disappeared, with only a plural marker remaining.

In this section, we will give a short overview of morphological processes and phenomena of Spanish, and briefly describe orthographical issues. We will present them in a way that allows us to define principles for the implementation of SMM.

A. Derivation

Verbs, adjectives, and nouns can form the base of a derivation. The base for derivation can be a simple word as well as a compound. Derivation happens through suffixes, prefixes, or a combination of both. Only suffixes can change the word class. Some suffixes require the insertion of a preceding interfix. Derivation includes conversion; e.g., participles of verbs can be used as adjectives.

Multiple derivations are possible, e.g., *inutilizable* 'unusable' is derived from the adjective stem *util* by adding the negating prefix *in*, the verbalization suffix *iza*, and the adjectivization suffix *ble*. In many such cases, the exact bracketing is debatable, e.g., whether the prefix was added to the result of the suffixation (*in+utilizable* 'un+usable') or whether suffixes were added to the result of the prefixation (*inutil+izable* 'unus+able'). Since there is no way for a morphological analyzer to determine the "correct" bracketing, it should thus *keep ambiguity with respect to bracketing and return it in a way that allows subsequent applications to resolve it.*

B. Compounding

Compounding – in the sense of combining free morphemes or well-formed word forms to form new words – is not used in Spanish to the extent it is used in languages like German. Compounds can be written as one word form (*sordo + mudo → sordomudo* 'deaf-mute'), with hyphens (*actor-cantante* 'singer-actor'), or as separate word forms (*treinta y uno* 'thirty-one'). Compounds written as separate word forms cannot be recognized by a morphological analyzer examining one word form at a time, but only by a tagger or during syntactical analysis.

Most compounds in Spanish are nouns or adjectives. Compounds can be constructed from nouns, adjectives, adverbs, and verbs. It is not always possible to unambiguously determine the resulting part of speech (POS). The principle is thus to *keep ambiguity with respect to POS and return it in a way that allows subsequent applications to resolve it.*

C. Inflection

Spanish word classes can be categorized into inflected classes (adjectives, nouns, determiners, pronouns, and verbs) and uninflected classes (adverbs, prepositions, conjunctions, interjections). There are two basic types of inflections, noun inflection and verb inflection. Only suffixes are used in inflection.

1) Noun Inflection: The principles of noun inflection apply to nouns, adjectives, determiners, pronouns, and numerals. For nouns and adjectives, gender and number are marked in the surface of word forms. Case is not marked and can therefore only be determined during syntactical analysis.

Pronouns and adverbs share many forms: When looking at an isolated word form it is not always possible to decide whether it is used as a pronoun or an adverb; we therefore assign the POS *Pronoun/Adverb*, and the analysis includes all information for both the pronominal and the adverbial readings. The final decision can only be made during syntactic or semantic analysis. Thus, as in the case of compounding, an implementation should *keep ambiguity with respect to POS and return it in a way that allows subsequent applications to resolve it.*

As the feature structures of Malaga are not restricted to a certain number of features or a certain structure of values, we propose to *gather as much information as possible during the analysis process*. If some of this information is not needed or wanted for a certain purpose it can easily be filtered out, which is much cheaper than trying to infer missing information.

2) Verb Inflection: In contrast to nouns and adjectives, the verbal inflection system is very rich. There are 17 possible combinations of mood and tense [3]; as verb forms are also marked for person and number, there are 111 word forms for each verb. However, some of these word forms share the same surface, so that it is not always possible to determine the exact category from the surface of an isolated word form. For example, the word form *cantara* (of *cantar* 'to sing') can be first *and* third person singular subjunctive imperfect. We therefore use the approach of *distinctive categorization* [1, pp. 244, 346]: Instead of postulating different word forms which are indistinguishable at the surface level, we only assume *one* word form which can have different functions. This drastically reduces the number of surface forms per verb to 52 – which still is very high when compared to English.

Spanish has three main conjugation classes, distinguished by the *theme vowel* (*a*, *e* or *i*) in each form of a verb. The information for person and number is marked using a single morpheme, and tense and mood are also indicated by a single morpheme. Traditional grammars (e.g., [4], [5]) thus arrive at the following segmentation for the word form *cantábamos* (first person plural indicative imperfect):

> *cant* (stem) + *a* (theme vowel) + *ba* (mood/tense) + *mos* (person/number)

However, as the combination of allomorphs for theme vowel, mood/tense, and person/number results in distinct strings for each combination, *cantábamos* can also be analyzed as

> *cant* (stem) + *abamos* (inflectional ending)

The ending is thus considered to contain all inflectional information. The ARIES system [6], [7] takes the same approach.

This leads to a further principle for the implementation: *Treat verbal inflection as concatenation of stem allomorph and inflectional allomorph. The inflectional morphemes yield all categorial information that will be presented in the result of an analysis.*

Traditionally, Spanish verbs are categorized as either *regular* or *irregular*. Irregular Spanish verbs exhibit irregularity only in the stem, the inflectional suffixes remain the same. Irregularities in the verbal stem may concern vowels only, consonants only, or vowels and consonants.

In fact, however, most of the "irregular" variation still follows certain rules. We therefore distinguish between *regular* (no variation), *semi-regular* (the variation can be derived from the surface of the base form), *semi-irregular* (the variation must be derived from a special surface marker), and *irregular* (suppletive) verbs, following the classification of Hausser [1, p. 263].

D. Orthographic Characteristics

Some orthographic conventions affect morphological analysis and generation. One case are accents. Spanish has certain fundamental stress patterns, e.g., word forms ending in a vowel, in *n*, or in *s* have penultimate stress. These cases are unmarked in Spanish orthography. If stress differs, the stressed syllable is marked by an acute accent (e.g., *derivación* 'derivation'). If phonologically legal, stress remains on the original syllable even if the number of syllables changes due to morphological processes. It can therefore be necessary to add, move, or remove an accent.

A similar case are phonemes that are represented by different allographs depending on the following vowel, e.g., /g/ is written as *g* before *a*, *o*, and *u*, and as *gu* before *e* and *i*.

While these are, strictly speaking, orthographic phenomena, it is necessary to handle them in a morphological component. This leads to a further principle: *Treat orthographic variants as allomorphs*.

E. Clitical Pronouns

Spanish is a language using clitical pronouns (*pronombres átonos*). Up to three clitical pronouns can follow a verb, e.g., *¡búsquesemelo!* 'find (pl.) it for me!'. Clitical pronouns can represent direct and indirect objects; the reflexive pronouns can also be used clitically.

It is debatable whether this is a morphological or a syntactical phenomenon: If a noun phrase is used instead of a clitical pronoun, the verb and the object are written as separate word forms, and thus do not appear to be a single word form. As an example, compare *¡dámelo!* 'give it to me!' to *¡da el libro a Maria!*[4] 'give the book to Maria!'; here, *me* is replaced by *a Maria* and *lo* is replaced by *el libro*.[5] If noun phrases are used, it is obviously the task of a syntactical component to check whether the phrase *el libro* is a valid valency slot filler for *¡da!*.

However, convention requires that a verb (in certain forms) and following pronouns are written without intervening spaces, thus giving the impression of being a single word form. The resulting "word form", though, is *not* part of the paradigm of the verb, as it results from neither derivation, nor compounding, nor inflection. We thus postulate a further principle: *The analysis of verb forms with clitical pronouns has to make clear that the surface consists of more than one word: The verb and the clitical pronouns.*

IV. Implementation of SMM

A. Principles and Approach

In section III we formulated several requirements we wanted our implementation to meet. To summarize, these are:

- Keep ambiguity with respect to bracketing and return it in a way that allows subsequent applications to resolve it.
- Gather as much information as possible during the analysis process.
- Use distinctive categorization.
- Treat verb inflection as concatenation of stem allomorph and inflectional allomorph.
- Distinguish regular, semi-regular, semi-irregular and irregular inflection.
- Treat orthographic variants as allomorphs.
- The analysis of verb forms with clitical pronouns has to make clear that the surface consists of more than one word: The verb and the clitical pronouns.

Furthermore, there are often several possible segmentations into allomorphs for a word form, all morphologically legal, but only some are likely to be semantically or conventionally acceptable. As a general principle, when ambiguity cannot be resolved on the level of morphology, a morphological component should not attempt to resolve it, as it could only guess. Instead, it should gather and return all relevant information, so that a higher-level component can use it to make a decision.

As noted above, SMM is based on the *allomorph approach* to morphological analysis: During analysis, word forms are segmented into allomorphs, which are then looked up in an allomorph lexicon; concatenation rules then combine the allomorph entries from the lexicon to determine lemma and category of the word form. The allomorph lexicon is generated (compiled) from a morpheme (base form) lexicon before run time; this means that during run time, only computationally cheap matching and concatenation operations are necessary. In contrast to systems based on full-form lexicons, the allomorph approach allows to analyze ad-hoc creations and neologisms, since the rules reflect the morphological processes of the language, and it only requires a relatively small base form lexicon.

SMM thus consists of three components: (1) The base form lexicon, (2) rules for creating allomorphs from these base forms (allomorphy rules), and (3) rules for concatenating these allomorphs at run time.

[4]Note that the accent on the verb is removed.

[5]Syntactical rules require different word order.

B. The Base Form Lexicon

The SMM base form lexicon contains 98,545 entries (see table I).

TABLE I
COMPOSITION OF THE SMM BASE FORM LEXICON

POS	# entries
Nouns	57,882
Adjectives	21,867
Verbs	12,826
Adverbs	2,517
Names	1,030
Acronyms	537
Interjections	317
Pronouns	89
Affixes	1130
Inflectional morphemes	126
Other	214
Total	98,545

Listing 1 shows three entries from the base form lexicon; *comer* 'to eat' is a regular verb, *huir* 'to flee' is a semi-regular verb (no markers required), and *decir* 'to speak' is a semi-irregular verb (the `AlloMark` feature contains the surface marker and the `AlloForm` feature indicates the applicable allomorphy rule).

```
[Lemma: "comer",
 POS: Verb,
 Valencies: <Reflexive, Intransitive, Transitive>];
[Lemma: "huir",
 POS: Verb,
 Valencies: <Reflexive, Intransitive>];
[Lemma: "decir",
 POS: Verb,
 Valencies: <Reflexive, Transitive>,
 AlloMark: "d{ec}", AlloForm: Allo_Norm_ecir1,
 P_imp_Sg2: <"di">, Participle: <"dicho">];
```

Listing 1. Entries for *comer*, *volver* and *decir* in the base form lexicon

An allomorph lexicon of 168,392 entries is generated from the base form lexicon by applying *allomorphy rules*, which take lexicon entries as input and create entries for the allomorph lexicon. The compilation of the allomorph lexicon takes about 9 seconds on an Apple MacBook[6].

The ratio of allomorphs per base form in SMM is 1.709. This is much higher than the ratio observed in other morphological systems implemented with Malaga [1, p. 268]. However, a large portion of the allomorphs differ only in the presence of a diacritical accent or due to orthographic rules as described in section III-D. Treating these variants as allomorphs is in line with other systems [6], [8], [9], [7] and allows for uniform processing, but the side effect is a high allomorphy quotient.

C. The Allomorphy Rules

The entries of the allomorph lexicon contain many more features than the original base form entries. The features from the base form entries are copied to the allomorph entries. Based on the POS and the surface specified in the base form entry, the allomorphy rules deduce certain features such as the theme vowel and the conjugation type for verbs.

[6]2.16 GHz Intel Core 2 Duo processor, 2 GB RAM, running Mac OS X 10.5.5.

```
"com": [POS: Verb,
        Valencies: <Reflexive, Intransitive,
                    Transitive>,
        BaseForm: "comer",
        ThemeVowel: e,
        PossibleEnclitics: 2,
        Suc: <<<POS, ThemeVowel>>,
              <<POS, Suffix>>,
              <<POS, Interfix>>,
              <<POS, VerbInflection>,
               <Tempus, <Inf, P_imp_Pl2, ...>>>>,
        SucFon: aeiou,
        Pre: <<<POS, Prefix>>,
              <<POS, Adverb>>,
              <<POS, Substantive|Adjective>,
               <WellFormed, yes>>,
              <<LastPOS, Punctuation>>>,
        Conjugation: regular]
"cóm": [POS: Verb,
        Valencies: <Reflexive, Intransitive,
                    Transitive>,
        BaseForm: "comer",
        ThemeVowel: e,
        PossibleEnclitics: 2,
        Suc: <<<POS, ThemeVowel>>,
              <<POS, Suffix>>,
              <<POS, Interfix>>,
              <<POS, VerbInflection>,
               <Allo_i, <encl1>>,
               <Tempus, <P_imp_Sg2, ...>>>>,
        SucFon: aeiou,
        Pre: <<<POS, Prefix>>,
              <<POS, Adverb>>,
              <<POS, Substantive|Adjective>,
               <WellFormed, yes>>,
              <<LastPOS, Punctuation>>>,
        Conjugation: regular]
```

Listing 2. Entries for *comer* in the allomorph lexicon.

Listing 2 shows the entries in the allomorph lexicon (with some feature values omitted) generated by allomorphy rules from the base form entry for *comer* (see listing 1). For the reasons outlined in section III-D, there are two allomorphs (*com* and *cóm*), even though *comer* is a regular verb.

All allomorph lexicon entries contain the two features `Pre` and `Suc`. `Pre` contains a list of features the preceding concatenated allomorphs have to have; `Suc` contains a list of features a following allomorph has to have. The `Pre` and `Suc` features essentially represent the morphological processes (inflection, derivation, compounding) and constraints on the level of allomorphs. For example, stem allomorphs of verbs include `<POS, VerbInflection>` in their `Suc` feature list, so that they can be followed by a verbal inflectional allomorph; verbal inflectional allomorphs, on the other hand, include `<POS, Verb>` in their `Pre` feature list. Similarly, prefixes cannot be followed by suffixes, etc.

The `Pre` and `Suc` features can also describe more specific constraints; see, e.g., the last `Suc` entry for *cóm* in listing 2,

which expresses that this allomorph is only used for imperative forms followed by clitics (e.g., *¡cómelo!* ‘eat it!’ vs. *¡come!* ‘eat!’).

D. Concatenation Rules

At run time, a single rule is used for concatenating allomorphs. The `Concat` rule takes two feature structures as input: *S*, the result of the analysis so far (including the concatenation of the allomorphs), and *N*, the lexicon entry for the next allomorph.

The rule processes *S* and *N* to produce a resulting feature structure S' and recursively calls itself with S' as value of *S* and N', the subsequent allomorph, as the value of *N*. When the input is exhausted, the special rule `FinalStateCheck` is called, which checks whether *S* represents a well-formed word form and defines the output, i.e., which of the features collected during concatenation will be shown to the user.

Our approach makes extensive use of information stored in the entries of the allomorph lexicon: The main principle is to check whether the properties of the next allomorph meet the requirements specified in the `Suc` feature of *S*, and whether the properties of *S* meet the requirements specified in the `Pre` feature of *N*. If this is the case, the features of S' are constructed from the features of *S* and *N*; S' does not need a `Pre` feature; the value of the `Suc` feature of S' is copied from *N*.

Thus, the main work is done in the allomorphy rules (1762 lines of Malaga code), which ensure that as much and as precise information as possible is available in the allomorph entries, whereas `Concat` and `FinalStateCheck` together amount to only 236 lines of code.

E. Analysis Results

Figure 1 shows the graphical output for the analysis of *volvemos* ‘we return’. Malaga allows to display all information accumulated during the analysis process. We decided to include the values for the features `POS`, `Baseform`, `Segmentation` (i.e., the allomorphs and morphological processes involved in concatenation), and `WordStructure` (i.e., information on the constituent morphemes) in the output. Depending on the word class, categorial and inflectional information is also included. All information can be stored and displayed in a structured way; this makes it easy for humans to assess results at a glance, and for programs to access specific information needed for further processing[7].

Figure 2 shows the graphical output for the analysis of *!dámelo!* ‘give it to me!’: a verb form with two clitical pronouns (see section III-E). The result is a list consisting of the analysis for the verb and a list with analyses for the pronouns. The analysis of the verb contains a marker (the `NextWord` feature) that one or more clitical pronouns are

"volvemos"

1:

POS:	Verb
ThemeVowel:	e
Conjugation:	semi_regular
WordForm:	"volvemos"
Segmentation:	"volv<FLX>emos"
WordStructure:	⟨ [Morpheme: "volver", Allomorph: "volv", POS: Verb], [Morpheme: "emos", Allomorph: "emos", POS: VerbInflection] ⟩
BaseForm:	"volver"
Valencies:	⟨Reflexive, Intransitive⟩
InflectionStructure:	⟨TV&PN⟩
Category:	⟨ [Tense: Present, Mood: Indicative, Person&Number: Pl1] ⟩
WellFormed:	yes
Status:	parsed

Fig. 1. Analysis of *volvemos*

following, and the analyses of the pronouns indicate that they are used as clitical pronouns (in the `PronounType` feature).

The hierarchical format of the analysis allows to individually access the information on the involved word forms, i.e., a syntactical parser has access to the verb and its features as well as to the pronouns and their features, as if they had been separate word forms in the input.

Malaga offers both graphical and text-based output formats, the latter being more suitable for further processing by scripts. For embedded use in other applications, Malaga provides an API.

As Zielinski and Simon note, “linguists generally are not only interested in the segmentation of a complex word, but also in its internal hierarchic structure.” [10] Unlike other morphologic analyzers, SMM has provided detailed, structured information from the very beginning.

F. Generation of Paradigms

Many applications require the capability to generate all word forms of a word (i.e., its paradigm) or the word form corresponding to a specific category. However, in contrast to systems based on transducers, it is not possible to simply “reverse” the analysis rules of a Malaga grammar. There are, however, two ways to work around this restriction.[8]

The first way is to write a separate “generation grammar,” i.e., a grammar that takes a base form and the desired category as input and returns the corresponding word form. For example, for an input such as `volver Verb Pl1 Present Indicative`, the grammar would return `volvemos`.

Since SMM encodes most of the combinatorial information in the allomorph lexicon (cf. the `Pre` and `Suc` features described in section IV-D), the effort for such a “generation grammar” is relatively low: The resources (lexicon and allomorphy rules) can be reused, and the concatenation rules

[7] We will outline programming interfaces and some potential applications in section VII.

[8] As only the verbal inflection is complex, the following discussion concentrates on the generation of verb paradigms.

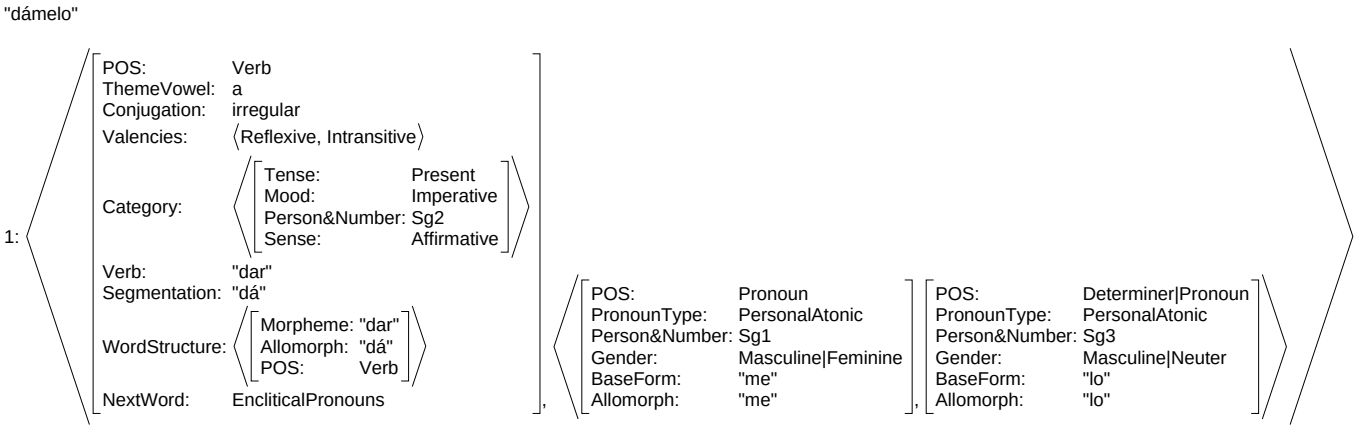

Fig. 2. Analysis of *¡dámelo!*

of the generation grammar primarily consist of mechanical agreement checks.

The second approach utilizes the `mg` function of Malaga. This function takes a list of allomorphs and a number indicating the maximum number of concatenated allomorphs as arguments. As result, the function returns all word forms that can be constructed from the given allomorphs, up to the indicated length. However, the result may include word forms not belonging to the paradigm, so it is necessary to filter the results, e.g., by using a small Perl script.

As we have described in section III-C2, we treat verbal inflection as the concatenation of a lexical stem and an inflectional morpheme (combining theme vowel, mood, tense, person and number). To generate the paradigm of a verb we can thus call `mg` with all allomorphs of the verb and all inflectional allomorphs and a maximum length of 2.

V. Performance and Evaluation

To measure the analysis speed and to get an impression of the performance we morphologically analyzed two corpora using SMM: The CRATER corpus[9] and a home-grown Web corpus (called WaC – Web as Corpus).

The CRATER corpus is a parallel English, Spanish, and French corpus consisting of ITU (International Telecommunications Union) documents. We used the manually tagged Spanish part and excluded all multi-word terms, foreign words, and non-word tokens[10] The WaC corpus was constructed by the method described by Sharoff [12], using the 500 most frequent Spanish word forms[11] as "seed words." Table II shows the detailed results for both corpora (WF: word forms). The performance data was collected on a Linux system[12].

The unrecognized word forms in the CRATER corpus are mostly typos and mistagged tokens. The recognition rate for WaC is lower, as the Web texts contain more typos, unmarked foreign words, non-standard abbreviations used in blogs and forums, etc. With respect to unique word forms, the recognition rate for WaC is extremely low. Furthermore, the difference between the recognition rates for WaC with respect to running word forms and unique word forms is much higher than the respective numbers for CRATER. However, most (86%) of the unrecognized word forms in WaC appear less than 5 times. Excluding these low-frequency unknown tokens increases the recognition rate to 93.2% for running word forms and to 84.7% for unique word forms.

VI. Related Work

There exist a number of other systems for automatic morphological analysis of Spanish word forms.

ARIES is a set of tools developed at the Universidad Politécnica de Madrid (UPM) [9], [7]. The morphological analyzer concentrates on inflection. From a base form lexicon of 38,000 entries 465,000 inflected forms are created using allomorphy rules. ARIES seems to be no longer maintained.[13]

COES [8], [13] is being developed at the UPM and the Universidad Carlos III. The COES tools are based on a lexicon of about 50,000 words, handle inflection, enclitic pronouns, and some types of derivation, but, as they are intended for spell-checking, do not provide analyses. COES has been integrated with ispell and is available under the GPL since 1994.[14]

AGME [14] was developed at the Instituto Politécnico Nacional, Mexico. It can be used for morphologic analysis and

[9]Corpus Resources And Terminology ExtRaction, see http://www.lllf.uam.es/ESP/proyectos/crater.html [last access 2009-02-04].

[10]Such as acronyms and numbers. Most of these tags are not listed in the documentation [11].

[11]Created from the list of the 1000 most frequent Spanish word forms from the *Corpus de referencia del español actual* of the Real Academia Española, http://corpus.rae.es/frec/1000_formas.TXT [last access 2009-02-04].

[12]2.2 GHz Dual-Core AMD Opteron Processor, 8 GB RAM, running Ubuntu 8.04 for x86-64.

[13]See http://www.mat.upm.es/~aries/ [last access 2009-02-04].

[14]See http://www.datsi.fi.upm.es/~coes/ [last access 2009-02-04].

TABLE II
SMM PERFORMANCE ON TWO CORPORA. (U) STANDS FOR UNIQUE WORD FORMS

Corpus	Word forms	Recognized	Results/WF	WF/s	Run Time
CRATER	422,953	417,481 (98.71%)	4.291	1,084	6 min 30 s
CRATER (u)	10,653	10,325 (96.92%)	8.821	409	26 s
WaC	125,685,532	116,128,864 (92.36%)	3.973	1,166	30 h 14 min 27 s
WaC (u)	1,576,530	671,258 (42.58%)	11.91	370	1 h 10 min 55 s

generation, but apparently handles only inflection. The lexicon consists of allomorphs with markers for which inflected forms they are used. The basis for this lexicon are 25,000 head words, from which 1,010,020 word forms can be generated. AGME uses an approach called *analysis through generation*: For analyzing inflected word forms several hypotheses for the category and the stem are produced. Then the generation is called with the hypothetic stems and categories and the resulting word forms are compared with the word form to be analyzed. The system can be downloaded in the form of a Windows executable.[15]

FreeLing [15] is an open-source suite of language tools, including a morphological analyzer, developed at the Universitat Politècnica de Catalunya. The Spanish dictionary contains about 550,000 word forms for 76,000 lemmas. [16]

The systems mentioned above differ in various respects, including their basic approach, lexicon size, and coverage of morphological processes. To our knowledge, there is no comprehensive evaluation of morphological analyzers for Spanish, so that currently analysis speed, coverage, and correctness of the systems cannot be assessed and compared objectively. The evaluation of morphological components is further complicated by the fact that different applications have different requirements (e.g., spell-checking vs. full syntactic analysis). Nevertheless, it can be said that SMM is set apart from other systems by its detailed, structured analysis results, whereas many other systems only provide a single tag. Furthermore, SMM handles *all* morphological processes of Spanish. Due to the allomorph approach, SMM is capable of analyzing previously unseen word forms created by derivation or compounding.

VII. USE IN APPLICATIONS

To make practical use of morphological analysis results in applications, it is critical that applications can integrate the morphological component and receive the results in a format suitable for further processing. Malaga provides a C library and API for this purpose and there are modules for Perl, Ruby, and Python, which allow convenient processing of analysis results.

Higher-level NLP applications which crucially require access to morphological information include syntactical analysis, semantical annotation of corpora, or rule-based machine translation.

[15]See http://www.cic.ipn.mx/~sidorov/agme/ [last access 2009-02-04].
[16]See http://garraf.epsevg.upc.es/freeling/ [last access 2009-02-04].

If we widen the focus to include "real world" applications, we find several scenarios in which morphological analysis is required. Information retrieval is a high-profile application which can benefit from morphological analysis. Another area is software for language learning; morphological analysis can be used to support instructors and students (e.g., by automatically extracting vocabulary lists from texts, by creating verb paradigms, or by automatically processing spelling, vocabulary, or grammar tests).

Language-aware functions for word processors [16] are another interesting field. Language-aware functions go beyond the services offered by spelling checkers and could support authors in tasks such as changing the tense, pluralizing expressions, or making global replacements sensitive to linguistic properties. Since these functions are used interactively, morphological analysis also needs to be fast.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented SMM, a morphological component for Spanish. It is implemented using the Malaga framework for developing grammars based on the formalism of Left-Associative Grammar. SMM handles all morphological processes (inflection, derivation, and compounding) and clitical pronouns. The design of SMM was influenced by the specific phenomena of Spanish, as well as by general principles, e.g., distinctive categorization. The Malaga framework allows hierarchically structured output suitable for further processing via programming interfaces. Thus, SMM is a component that can be easily integrated into batch and interactive applications. Using the internal functions of Malaga, it is also possible to use SMM for generating verb paradigms and specific word forms.

A future task will be the introduction of weighting to reduce the number of analyses. Weighting offers the possibility of having SMM return either only the most probable analysis or all analyses, ranked according to their probabilities.

For evaluation, we are considering to compare the SMM analyses with the manually tagged CRATER corpus. This will require a mapping of SMM categories to CRATER tags. Unfortunately, preliminary tests have revealed a non-negligible number of tagging errors and have shown that actual usage of the tags differs from the documentation [11], so that further analyses will be necessary.

SMM will be published as open-source software in the next months.

ACKNOWLEDGMENTS

We thank Björn Beutel for the development and maintenance of Malaga.

REFERENCES

[1] R. Hausser, *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*, 2nd ed. Berlin/Heidelberg: Springer, 2001.

[2] ——, "Three principled methods of automatic word form recognition." in *VEXTAL: Proceedings of the Conference, 22–24 November 1999, Venice, Italy.* Padova: Unipress, 1999, pp. 91–100.

[3] Real Academia Española Comisión de Gramática, *Esbozo de una nueva gramática de la lengua española*, 2nd ed. Madrid: Espasa Calpe, 1974.

[4] J. Alcina Franch and J. Manuel Blecua, *Gramatica española*, 9th ed. Barcelona: Ariel, 1994.

[5] I. Bosque and V. Demonte, Eds., *Gramatica descriptiva de la lengua española*. Madrid: Real Academia Española/Espasa Calpe, 1999.

[6] J. M. Goñi Menoyo and J. C. González Cristóbal, "A framework for lexical representation," in *AI95: Fifteenth International Conference. Language Engineering*, June 1995, pp. 243–252.

[7] A. Moreno-Sandoval and J. M. Goñi Menoyo, "Spanish inflectional morphology in DATR," *Journal of Logic, Language and Information*, vol. 11, no. 1, pp. 79–105, 2002.

[8] S. Rodríguez and J. Carretero, "A formal approach to Spanish morphology: the COES tools," in *XII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'96)*, 1996, pp. 118–126.

[9] J. M. Goñi Menoyo, J. C. González Cristóbal, and A. Moreno, "ARIES: A lexical platform for engineering Spanish processing tools," *Nat. Lang. Eng.*, vol. 3, no. 4, pp. 317–345, 1997.

[10] A. Zielinski and C. Simon, "Morphisto: An open-source morphological analyzer for German," in *Seventh International Workshop on Finite-State Methods and Natural Language Processing*, 2008, pp. 177–184.

[11] F. Sánchez León, "A Spanish tagset for the CRATER project," Jun 1994. [Online]. Available: http://arxiv.org/abs/cmp-lg/9406023v1

[12] S. Sharoff, "Creating general-purpose corpora using automated search engine queries," in *Wacky! Working Papers on the Web as Corpus*, M. Baroni and S. Bernardini, Eds. Bologna: GEDIT, 2006.

[13] S. Rodríguez and J. Carretero, "Formalización de reglas morfológicas para un nuevo corrector ortográfico en español," *Revista Española de lingüística*, vol. 26, no. 2, pp. 379–387, November 1996.

[14] A. Gelbukh and G. Sidorov, "Approach to construction of automatic morphological analysis systems for inflective languages with little effort," in *Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing 2003, Mexico City, Mexico, February 16–22, 2003*. Berlin/Heidelberg: Springer, 2003, pp. 157–162.

[15] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró, "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, 2006, pp. 48–55.

[16] C. Mahlow and M. Piotrowski, "Linguistic support for revising and editing," in *Computational Linguistics and Intelligent Text Processing: 9th International Conference, CICLing 2008, Haifa, Israel, February 17–23, 2008*, A. Gelbukh, Ed. Berlin/Heidelberg: Springer, 2008, pp. 631–642.