Language, Information and Computation(PACLIC 11),1996, 425-437

Spanish Lexical Acquisition via Morpho-Semantic Constructive Derivational Morphology

Margarita C. González

Psychology Department,
Computer Science Department,
New Mexico State University,
Las Cruces, NM 88001, USA
mgonzale@crl.nmsu.edu

Abstract

This paper describes an algorithm for Spanish derivational morphology whose output is generalizable to two different lexicon acquisition situations. One is the process of automatic lexicon acquisition via the use of Morpho-Semantic Lexical Rules (MSLRs), (Viegas, González, & Longwell 1996) usable in semantically based Natural Language Processing(Nirenburg, et al 1996) in order to considerably reduce acquisition time per entry in a large-scale semi-automatic acquisition environment (Viegas, Onyshkevych, Raskin, & Nirenburg 1966b). The other is its application in the language classroom to fascilitate vocabulary acquisition for second language learning students. It is an ancillary tool to be used in reading or writing tasks. The constructive approach used in the design of this tool addresses the overlap between the morphological and semantic criteria while implicitly addressing morpho-phonological phenomena. The objective is accomplished by means of three major strategies: pre-categorized base stems, inherited stem changes, and eight separately identified patterns of attachment which produce noun, adjective, verb and adverb subsets. These mechanisms produce words depending on whether the word structure is right-headed or left-headed, based on the type of affix producing Morpho-Semantic Lexical Rules (MSLRs) used and on the sequence in which affix attachments succeed. The set of MSLRs, approximately 151, produces words with their respective semantic component for the morpheme or sequence of morphemes, thus defining each derivational paradigm. Their product includes not only simple and complex affixation but also word compounding, allowing also overgeneration of derived forms for broader coverage of written and oral word forms. Unnecessary overgeneration of output paradigms can be automatically checked against the contents of machine readable dictionaries and against electronic texts. The tool includes a Graphical User Interface (GUI) in order to facilitate the vocabulary acquisition process for the L2 learner of Spanish.

1 The GEN_WORD Objective

A dual purpose guided the design process while taking advantage of a simple yet sophisticated and productive features of language in general.

- To automate the production and acquisition, via Morpho-Semantic Lexical Rules, of a large lexicon to be used by a machine translation system, using a computational semantic approach.
- To automatically produce data to be accessed by means of a GUI in order to facilitate the vocabulary aquisition process for Spanish as a Second Language learners.

2 GEN_WORD: System Description

GEN_WORD is an attempt to describe Spanish derivational morphology through constructive generation of derivational paradigms. It provides a morphologically derived Spanish lexicon. Each well-formed word has information concerning its syntactic structure and its semantic counterpart. The output format provides this information in the form of a derivational paradigm, usually generated from a verb stem. Thus, some back-generation for nouns and adjectives which themselves were derived from complex verb forms is possible.

This implementation is intended to provide broader coverage of well-formed words and their corresponding morpho-semantic descriptions.

GEN_WORD addresses all predictable morpho-phonological phenomena and the overlap of morpho-lexico-semantic restrictions via a set of three major components which constitute the derivational process. The words form using an affix lexicon which includes the corresponding morpho-semantic classification, procedural stem changes, and eight unique patterns of attachment.

2.1 Background on the Problem of Open-Class Derivation

Current approaches to derivational morphology have three issues. There is an ever present gap to bridge between theoretical approaches in design and practical, human-user oriented, applications. Program developers do not always consider learning about the users' actual needs; their main preocupation centers on computational efficiency, and output quality suffers as result. Moreover, a specific problem, inherent to open-class word derivation, is the treatment of both, overgeneration and undergeneration of well-formed words.

Because the usual intent is to present a formal description of morphology following a specific theoritical ventage point, these algorithms are not easy to use nor is their output all that usable in practical applications. Two-level PC morphology programs (Koskenniemi, 1983), parallel sets of transducers (Kay, 1970), and finite state automata (for Spanish, Martí 1986; Meya 1986; Tzoukermann and Liberman 1990, as cited by Moreno-Sandoval 1991; Jokinen (1993) for Finnish) are difficult to use since the design caters to a small group of users, well-versed in phonology and morphology. In addition, the focus of transduction rules for analysis and generation is on orthography and affixation rule, rather than on the description of an overall paradigm within- and between- inflectional and derivational paradigms and their semantic links which convey meaning.

There is, however, somewhat of an exception. Jokinen (1993) derives Finnish verbs and extends the formalism to include a virtual lexicon used for the categorization of well-formed words with their syntactic terminal elements. But she applies semantico-syntactic lexical rules separately from the morphological analysis, thus combining open-class derivational and inflectional analyses. Still, the inheritance structure typical of morpho-semantic lexical rules which can produce derived forms along with their paradigmatic characteristics is not used.

The problem is that program developers concentrate on the computational aspects of the problem and project what the needs of the user should be, rather than carrying out protocol analysis of the user engaged in the actual use of open-class derivational morphology. The thrust toward the preference of finite state machines (FSTS) lies in computational efficiency. Yet two-level morphology, especially FSTs, do not provide elegant treatments of reduplicative or template morphology (Sproat, 1992). Despite Goldman's (1993) attempt to use an interpreter in order to simplify two-level rules, it is not that clear whether his objective is achieved. His approach may even complicate matters altogether. As for the unification based approach by Moreno-Sandoval (1991), it is only useful for Spanish with respect to inflectionally derived word forms and their syntactic information. He succeeds at presenting synthesis and analysis via a single bidirectional method. Due to an extremely large lexicon with surface strings and grammatical categories, the approach is not easily generalizable to other languages, be they related or unrelated to Spanish.

The only other unique attempt at presenting a more linguistically reasonable descriptive model of morphology is Cahill's (1989) syllable-based MOLLUSC. This approach uses a tree-like structure common of the inheritance of phonological features in a morpheme. That is, it focuses on the construction of syllables, thus providing a microscopically segmented view of morphology. A greater disadvantage of MOLLUSC is the output of "phonological (semi-)realized forms".

Finally, equally important are computational efficiency and quality output. Quality output requires that inherent problems to derivational morphology be given proper consideration. That is, undergeneration of well-formed words and overgeneration of legal lexical forms (lexical gaps) must receive the attention they merit. Unfortunately, those tools that admit to some degree of overgeneration, do so to the extent that a nonsense-word, whose structure displays a composite of actual morphemes following phonotactic and graphotactic constrains, can be entered and a true word is the product of the derivation process(Antworth, 1990). Yet, if a perfectly legal word is not found in the dictionary being used by the algorithm, then that word, no matter how high its frequency rate, will not be part of the output.

More recently, work in computational morphology has moved to a statistical approach involving statistical probability. Thus, output contents are curtailed, being a clear case of undergeneration. The work of Sproat (1993) takes this approach. Again, the product is merely a lexicon of surface forms with part-of-speech tags. A similar tool is available for German (Lezius, Rapp & Wettler, 1996). However, it presents part-of-speech information but no morpho-semantic relations. Furthermore, depending on the size of the corpus to be tested, word recognition can vary from 95.0% to 81.8%. These percentages of word recognition may be reasonable for automatic word processing, but not for human users. When a word is needed to be analyzed, the tool must be able to process it or to present other options for dealing with the query. L2 students need to know whether a word is well-formed and whether or not it is usable in some context.

2.2 Componential Treatment – The Morphological Lexicon

Word forms are either lexical forms (found in the lexicon) or surface forms(found in the dictionary or in electronic texts). Lexical forms are the actual affix allomorphs of all bound morphemes in Spanish. On the other hand, surface forms are the actual words that serve as input to GEN_WORD; a stem lexicon is, therefore, irrelevant. There are no intermediate forms to be rewritten or repaired. These are implicitly inherited as the construction rules operate on the base stem taken from the surface form.

The system uses a considerably small lexicon (when compared to that used by Tzoukermann and Liberman (1990)) of affix allomorphs (2,259). These attach to base stems of the surface forms (verbs). All distributional criteria (assimilation, dissimilation, deletion, palatalization, and metathesis) applicable to the affix environment in Spanish word formation are implicitly represented. This representation is found in the set of allomorphs per morphemic description in the lexicon, in the dynamically produced variations of the verb stem and in the morpho-semantic attachment rules.

Affixes are categorized as either prefixes, infixes, or suffixes and are appropriately found in different modules. However, some prefixes and suffixes become infixes, and they do so implicitly. Here is the treatment of allomorph entries for the POLARITY_NEG prefix morpheme:

```
prefix1("i", polarity_neg1,[leg],[1a]).
prefix1("ir", polarity_neg2,[rom],['2B']).
prefix1("im", polarity_neg3,[per,pre],['1a']).
prefix1("in", polarity_neg4,[con,tra,fir],['1R','1S','1a','2I','2R','2a', '2b','2h','2k','2m','2p']).
```

This indicates that for all '1a' verbs whose stem-initial environment is of the form leg-, the allomorph i- of the POLARITY_NEG morpheme would be the corresponding attachment. Those verbs with '1R', '1S', '1a', '2I', '2R', '2a', '2b', '2h', '2k', '2m' and '2p' categories

whose stem-initial environments are con-, tra-, and fir-, correspondingly attach to the POLARITY_NEG4 allomorph in-.

As for words such as "acción", the morpheme -ión has several allomorphs. The following example includes the actual version of the output generated by the rules for the corrresponding entries in the lexicon.

ending1("ión",n, lr2event4a,lim,imp,['1b']). → limpión, n, lr2event4a

ending1("ción",n, lr2event4b,ad,str,['1a']). → administración, n, lr2event4b

ending1("sión", n, lr2event4c,exp,lot,['1a']). → explosión, n, lr2event4c

ending1("xión", n, lr2event4d,con,ect,['1a']). → conexión, n, lr2event4d

ending1("ón", n, lr2event4le,enc,ntr,['1l']). → encontrón, n, lr2event4e

Using the same rationale, there is a separate lexicon module for building compound words, thus facilitating compilation.

2.2.1 Procedural Stem Changes due to Verb Categorization

Once a surface form is processed and a category is assigned to the base stem, all base stems undergo an individualized stem change.

The stem classification was obtained as verb lists were preprocessed via a separate implementation (CAT_VERB). The latter is the result of the need to improve on the verb categories supplied by Collins Bilingual Dictionary, and the need to categorize by the same classification system all verbs found in other dictionaries as well as those verbs found in corpora alone. The verb category is arrived at by the analysis of the stem-final and/or stem-initial environments. (The resulting general category is used to process these verbs with either the inflectional (GEN_CONJ) or the derivational (GEN_WORD) implementation.)

GEN_WORD uses the specific category assignment to perform the appropriate stem changes. Subsequently, GEN_WORD uses the category and the analyzed particles of its stem-initial, stem-end distributions for the proper affix selection and application of construction rules. That is, GEN_WORD produces unique changes for all "1a" base stems. Thus, it separates these from any other base stems whose initial-, medial-, or final-stem environments may be similar but belong to other categories. This is so because of the existence of dual spelling verbs, as well as verbs that can have the same spelling in their infinitive forms but whose semantics are different and whose derived forms have different spelling. Consequently, this derivational implementation requires a general category and a sub-category to identify the method of processing a given base stem. This procedure generates at least one stem change and at most eight changes, depending on the base stem category. Verbs like *materializar* or *finalizar* will undergo, at most, two root changes:

materializ → materializa, materiali

The verb *restregar* undergoes changes for four variations, two of which have a stem-medial distribution:

restreg → restrega, restregu, restrieg, restriegu

The verb *celebrar* has a stem-initial syllable change that corresponds to a suprasegmental modification:

celebr → celebra, celebri, celeb, célebr.

Base stems are processed through the stem changing procedure only once, for the most part. Some post-positional affix attachment operations may call that procedure again when base stem category changes are the result of the derivational attachment (i.e., the verb *romper*). Each procedure and its attachment rules can operate on any of the changed stems as well as on the base stem. Invariably, all base stems need at least one stem-change. In Spanish, this is the unique characteristic of well-formed words derived from verbs. The vowel of the "-ar, -er, -ir" infinitive endings or an allomorphic variation thereof will attach to the stem and produce such words as:

cancelar → **cancel[a]do, cancel[a]ción,** etc.,

beber → **beb[e]dero, beb[e]dor**,...,but also

beb[i]do, beb[i]da, etc. ([i] is derived from the past participle of the verb.)

vivir → **viv[i]dor, viv[i]do,viv[i]enda, viv[i]ente**, etc.

The [a], [e], [i], are inherited from the -ar, -er, and -ir endings, respectively.

One advantage to having a dynamic procedure make the necessary changes in the base stem is to comply with the phonotactic (phonemic or suprasegmental) requirements of each affix attachment rule and its appropriate graphemic representation. Thus, this procedure eliminates the multiplication of allomorph entries in the prefix and suffix lexicons by n, where n = 1,...,6 possible stem changes per verb (17,000 verbs) to be processed. Therefore, GEN_WORD does not need to have each unique stem change as a separate entry in a stem lexicon (as seen in Tzoukerman & Liberman (1990)). The added advantage of this treatment is that it allows for the ommission of orthography rules at any given time within the attachment process. Such extra processes of modification, addition, and or deletion of graphemes are required of the syllable-based (Cahill, 1989) or rule-based morphology (Anick and Artemieff, 1992). The stem changing procedure is called only once, and its results are inherited by all its subsequent attachment procedures.

More importantly, graver processing problems common in Kimmo-based (Antworth, 1990; Karp et al., 1990; Ritchie, 1992; Karttunen, 1983; Karttunen and Wittenburg, 1983; Koskenniemi, 1983; Koskenniemi and Church, 1988) and FST-based (Martí, 1986; Meya, 1986; Tzoukerman and Liberman, 1990) software are also avoided. Unlike these implementations, GEN_WORD does not produce a perfectly legal word from a nonsense word input.

2.3 Componential Treatment – The Attachment Patterns

A natural tree-like inheritance hierarchy contributes all derivational patterns of well-formed words to the complete paradigm for each verb being processed. This process, in the very general sense of inheritance, is similar to Anick and Artemieff's Paradigm Description Language (1992). There is one important difference between GEN_WORD and PDL: PDL applies to French inflectional verb paradigms. GEN_WORD, on the other hand, is an open-class word derivational application for producing Spanish words belonging to different parts of speech.

A derivational pattern is a set of form construction rules which characterize a particular subset of surface forms belonging to the general paradigm. The tree-like inheritance works throughout each one of the patterns. Each form rule allows for only one process of affixation. This, in turn, allows the implicit inheritance of each partially derived form from the previous attachment process to which a subsequent attachment is made until the complete tree is traversed and the well-formed word is produced. No additional corretions or deletions are needed on the partially derived forms. This is the direct result of the implicit inheritance of the appropriate stem changes from the point at which the stem changes were produced. The stem-change procedure is called at the outset of a call to a set of patterns.

The default inheritance across patterns is universal. All patterns inherit the verb category, the verb stem, and the analyzed stem-initial and stem-final particles. However, each pattern has a unique set of characteristics with respect to the attachment types and their directionality of attachment. While prefix attachment rules require only the stem-initial environment information, the final process is always suffix attachment for all words. At this point all patterns will require the stem-final particle information, which varies according to the immediately previous attachment. Therefore, subsequent patterns to be produced cannot directly inherit what previous patterns have already generated. (Pattern inheritance would only be successful if GEN_WORD were concerned with inflectional morphology.)

An example of how the output would be negatively affected if pattern inheritance were implemented is presented with the word *similar* in the output paradigm for the verb *asimilar*. The autonomous pattern approach taken in GEN_WORD clearly generates the words *asimilar* and *desasimilar*, and it has potential to produce other words like *inasimilable* and *asimilabilidad*.

The paradigm output for patterns 1, 2 ,3, 4, and 5 of the verb *asimilar* is as follows:

similitud n lr2literal_attribute8b

similor n lr2literal_event_attribute1a

similar adj lr3reputation_attribute2f

similitudinario adj lr2attribute8a lr3attribute6 lr3reputation_attribute4e

asimilar v beside_spatial_relation1 lr1event

asimilación n lr2event4a

asimilista n lr2reputation_attribute9b

asimilable adj lr3feasibility_attribute1

asimilista adj lr3reputation_attribute9b

asimilativo adj lr3social_role_relation5c

asimilativamente adv lr3social_role_relation5d lr4object_relation1

desasimilar v opposite_relation2 beside_spatial_relation1 lr1event

desasimilación n lr2event4a

inasimilable adj ineg_coutner_relation1 lr3feasibility_attribute1

inasimilablemente adv lr3feasibility_attribute1 lr4object_relation1

asimilabilidad n beside_spatial_relation1 lr2attribute1 lr2feasibility_attribute10a

If the same set of rules for pattern_1 were to subsequently apply to pattern_3, none of the words in the subgroups for *asimilar, desasimilar, or inasimilable* would be generated. What is worse, incorrect variations of *similor* and *similitud* would take up all the prefixes, thus producing many non-words and undergenerating legal words.

Separate generation of patterns is the approach taken in this implementation, though all paradigms have access to the same prefix and affix lexicons. (See Appendix A for two examples of pattern attachment.) This is done in an effort to constrain generation with clear and definite inheritance hierarchies. Thus, non-word formation is avoided.

The pattern of attachment processes are organized in increasing order of complexity, following basic notions of language complexity with respect to language acquisition. As in the example for the verb *asimilar*, the verb must access patterns 1, 2, 3, 4, and 5. The verb is processed through each of these patterns consecutively but without reference to what has transpired in the previous pattern. This order also helps present implicit derivational inheritance in the output of the derivational paradigm. As seen above with the verb *asimilar*. *desasimilar* derives from *asimilar*. The verb *similar* does not exist. Similarly, *inasimilable and inasimilablemente* are all derived from *asimilable*.

Within patterns, the default inheritance is representative of the morpho-semantic complexity of affixation. It also represents the part-of-speech category set of words that each verb could possibly generate and can be prescribed apriori. Unlike any other bottom-up approach, be it finite state (Tzoukermann and Liberman, 1990; Kay, 1987; Martí, 1986; Meya, 1986) unification based grammars (Moreno-Sandoval, 1991), rule-based (Anick and Artemieff, 1992; Cahill, 1989), even two-level morphology (Antworth, 1990; Goldman, 1993; Karp et al., 1990; Karttunen, 1983; Karttunen and Wittenburg, 1983; Koskenniemi, 1983,

1984; Koskenniemi and Church, 1988; Ritchie, 1992; Ritchie et al., 1992), concerned with phonotactics and graphotactics with corresponding grammatical labels and part-of-speech alone, GEN_WORD is an integrated top-down/bottom-up approach. Although Jokinen attempts to define Finnish verb derivation via lexical rules and finite state automata, the formalism is based on the correspondence established between the feature sets of a stem and a well-formed word found in a virtual lexicon. Essentially, the lexical labels are added after the words are produced (Jokinen, 1993). Finally, GEN_WORD attempts to describe morpho-semantic complexity in derived word forms. While it is true that Russell et al. (1991) are also concerned with a "multiple default inheritance" for the description of lexical entries, their concern lies in the partitioning of syntactic, semantic, and morphological classes of behavior, not on examining the working inter-relationship of these classes, especially semantics and morphology.

In addition to stem_initial and stem_end segments, and the verb category, within- pattern processes add other information to the inheritance structure. Morpho-semantic, part-of-speech information pertaining to the affixes to be attached as well as the affixes themselves, graphotactic stem changes, and the implicit order of attachment constitute this information. This information is acquired as the attachment process ensues. Thus, the well-formed word output will have the proper morpho-semantic information and the part-of-speech. The morpho-semantic label, in turn, becomes the name of the lexical rule that applies to the derived form.

Within-pattern constraints also allow a certain amount of overgeneration of such well-formed, legal words which may be generated in order that well-formed words that actually exist either in written or oral form, though not normally found in dictionaries, can also be generated. For example, all nouns that are events are derived from a verb whose obvious morpho-semantic category is lr1event1. Derived forms stemming from the verb *comunicar* are such an example.

comunicar → telecomunicación, n, spatial_relation1 lr2event

But *telecomunicar* is generated before *telecomunicación* can be generated. That is, *telecomunicación* derives from *telecomunicar*.

Thus, the set of eight mechanisms follow unique paths for affix attachment, which use two important criteria: left- or right-headedness and the number of affixes (the number of words, in the case of word-compounding) to be attached. That is, for Spanish, prefixes will often determine whether or not there is a change in the syntactic and semantic categories of the word being derived. However, stem-post-positional affixes can also determine such a change. But in the case of both stem pre-positional and post-positional affixation occurring in the same derived form, there is an interaction between prefix and suffix so that the information of both provides the complete information for the morpho-lexico-semantic information that the well-formed word inherits.

2.4 Componential Treatment – Affix Selection and Attachment Rules

A unification procedure selects the affix from the lexicon, and a concatenation procedure makes the appropriate stem-affix or affix-stem attachment inherited by the output rules.

These rules (approximately 151) perform the specific attachments of phono-morphologically correct affixes in their graphemic transcription. The corresponding lexico-semantic rule, the combination of the attachment particles and the verb category serve as the index for attachment. The morpho-phonological information is represented in the form of allomorphs for each morpheme to be attached pre- or post-positional to the stem. The base stem, as analyzed from the input word, or one of its various stem changes, is attached to the appropriate affix found in the lexicon with respect to its morpho-lexico-semantic category. To help constrain overgeneration, a set of procedures blocks off certain derived forms and permits only the specified set of prefixes to be attached according to the morpho-semantic

category (in those patterns where prefix attachment occurs). In all patterns, attachment processes are constrained by the set of part-of-speech categories (verb, noun, adj, adverb, or any permutation of these) that can be generated from each base stem.

2.4.1 Lexical Rules

The inheritance structure used in this implementation promotes the construction of words having a Morphó-Semantic Lexical Rule (MSLR) per morphemic attachment or a set thereof when the attachment of more than one morpheme contributes information to the rule.

The lexical rules are used to project the semantic contributions of morphemic attachment. These derived features do not percolate since they have the possibility of being multivalued, but they can also be single-valued. The latter is exemplified by verb forms that have been nominalized. A case in point are the lexical rules LR2EVENT, LR2LOCATION, LR2ORGANIZATION. The attachment of the morpheme -ción, a right-headed attachment, adds a value to the features of the nonhead–the verb stem of *administrar*, thus producing:

administración n lr2event4a

administración n lr2location14a

administración n lr2organization1a

The argument structure of the head often acts as an operator of the non-head. The head is the suffix morpheme, the non-head is administr-. All these well-formed words inherit the arguments of the verb stem.

Similar examples are also common for adjectives and their derived adverbs.

Examining the derived, multi-valued, forms of the verb *romper*, the prefix ir-, NEGATIVE_COUNTER_RELATION5, overrides romp-, thus forcing a change in the stem to -rump- and forcasting the attachment of the lr1event suffix morpheme -ir as one derivational option. Thus, romper can become *irrumpir*, or *irrompible*.

The same argument can be used with compound words. the stem compr- produces the word compra, LR2_EVENT8B, which in turn serves as a head for **compraventa, n, lr2event8b lr2event8b**

Another issue closely linked to inheritance is the matter of the process by which affixes absorb an argument of their base (Booij and van Haaftern, 1988). The morpheme -iz creates agent and/or instrument nouns and binds the external argument to the verb stem. It is a causative affix that adds an agent role.

comerciar → comercial, n, lr2event26b

comercializar v lr1event

comercialización n lr2event4a

comercialización n lr2theme_of_event10a

comercializado adj lr3event_telic1a

comercializado adj lr3reputation_attribute1a

The following sentences demonstrate the change in meaning.
Las acciones son comerciales. administradores han comercializado las acciones.

2.5 Output rules

At present the output rules produce the well-formed word, its part-of-speech and morpho-semantic category, or lexical rule label. The necessary information for displaying derivation or analysis becomes available in each step of the derivational process. In this case, whether or not the program's implementation is characterized by bidirectionality is irrelevant.

One additional function realized on the output is to execute a search of the word forms on machine readable dictionaries or electronic corpora. If the word is found as a headword or cannonical form in the dictionary, the word and its label are deposited in one file. Otherwise, words not found and stored in residual lists can be checked against electronic corpora (written language or transcriptions of oral language) as the need arises.

2.6 GEN_WORD a tool for Spanish Second Language Learners

In the hopes of facilitating the vocabulary acquisition process for students of Spanish as a second language, a graphical user interface was designed as a front end for GEN_WORD. Prior to undertaking this design task, a considerable amount of time was invested in analyzing evidence from current research in applied linguistics and psycholinguistics. Topics concerning the cognitive processes in cross-linguistic transfer, lexical access and L2 morphological awareness in reading for bilinguals were the main focus. The GUI design emerged from this research and 20 years of actual teaching experience in the L2 classroom. (See Appendix B for a graphical representation of GEN_WORD's GUI.)

3 Conclusions and Future Concerns

The reported research concentrated on the use of lexical rules for derivational morphology as the driving mechanism for stem and affix unification within a larger process of blocking for producing paradigmatic output. Furthermore, lexical rules have been shown, in small-scale experiments, to work for other kinds of lexical regularities, notably cases of regular polysemy (Ostler and Atkins, 1992; Apresjan, 1974).

The treatment of transcategoriality supports the hypothesis that a hierarchical organization of derivational morphological behavior helps to explicitly organize the generated output of well-formed words into a hierarchy of a few "original" senses and a number of senses derived from them according to well-defined rules. Furthermore, the semantic inheritance of the base stem is also transparent. Thus, the argument between the sense-enumeration and sense-derivation schools in computational lexicography may be of less importance than suggested by recent literature.

If lexical rules designed for this project are compared to others used in lexically-based grammars (such as (GPSG, (Gazdar et al., 1985) or sign-based theories (Pollard and Sag's (1987) HPSG), the latter can be viewed as linking rules that often deal with the subcategorization issue. These rules were the driving force of unification mechanisms and paradigm design, as mentioned before, thus, producing overgeneration to a certain extent. However, overgeneration in this particular situation is considered an asset rather than a liability; the mechanisms employed by the formalism rely on linguistic clues which allow a limited amount of overgeneration. Furthermore, results from the evaluation of the Lexicon acquisition task for Mikrokosmos evidently indicate this approach used in large-scale automatic lexicon acquisition is useful. Despite the fact that 80% of the Spanish acquisition entries were generated by GEN_WORD, significant time and effort was spent on semi-automatic checking of the generated output and the corresponding MSLRs. The latter problem was resolved by automatically generating a reverse lexicon, whose regenerated entries were used to test the thousands of MSLR automatically created entries. (Viegas & Beale, 1996).

Secondly, the intent was to produce a morphological description tool for users whose level of sophistication in linguistic knowledge is minimal using the same output. Overgeneration in this case is also advantageous, thus providing the student with new possibilities to explore. The tool includes several categorizational schemes imposed on GEN_WORD's output. These allow the L2 student to view derived well-formed words and their morpho-semantic relation from a distinct perspective, be it a single-word, a complete paradigm, the different part-of-speech categories, suffix and prefix-suffix patterns of attachment, as well as compound words. Thus, GEN_WORD's facilitation of reading and writing exercises in Spanish as a Second Language can be supportive of the vocabulary acquisition process.

Finally, improvements to GEN_WORD are forthcoming in the following areas: morphological rule enhancement to take into account the Spanish verb transitivity/ intransitivity; generalizability to other languages, especially within the family of Romance Languages; integration of both syntactic/semantic morphological generation and analysis (integration of GEN_CONJ and GEN_WORD in a general tool);and GUI design enhancement adding a graphic tree inheritance implementation for single words and for word paradigms.

4 References

Anick, Peter & S. Artemieff. 1992. A high-level Morphological Description Language Exploiting Inflectional Paradigms. Actes de Coling-92, 23-28. Nantes, France: COLING-92.

Antworth, Evan L. 1990. PC-KIMMO: A Two-level processor for Morphological Analysis. Occasional Publications in Academic Computing 16. Dallas , TX: Summer Insitute of Linguistics.

Apresjan, Ju. D. 1974. Regular Polysemy. Linguistics 142: 5-32.

Booij, Geert & T. van Haaften. 1988. On the External Syntax of Derived Words: Evidence from Dutch. Yearbook of Morphology 1:29-44.

Cahill, Lynne J. 1989. Syllabe-based Morphology for Natural Language Processing. Doctoral Dissertation, University of Sussex England.

Gazdar, Gerald, E. Klein, G. Pullum, & I. Sag. 1985. Generalized Phrase Structure Grammar. Blackwell: Oxford.

Goldman, Robert P. 1993. Direct implementation of two-level rules in prolog. Unpublished. New Orleans, LA: Tulane University.

Jokinen, Kristiina. 1993. Lexical Rules and Finite State Automata. Unpbulsined, Helsinki, Finland: University of Helsinki.

Karp, K., Y. Schabes, M. Zaidel, & D. Egedi. 1990. A Freely Available Wide Coverage Morphological Analyzer for English. Philadelphia, PA: Department of Computer and Information Science, University of Pennsylvania.

Karttunen, Laurie. 1983. KIMMO: A two-level morphological analyzer. *Texas Linguistic Forum*, 22: 165-186.

Karttunen, Laurie & K. Wittenburg. 1983. A two-level morphological analysis of English. Texas Linguistic Forum, 22:217-228.

Kay, Martin. 1987. Nonconcatenative finite-state morphology, ACL Proceedings, 3rd European Meeting.

Koskenniemi, Kimmo. 1983. Two-level Morphology: A General Computational Model for Word-form Recognition and Production. Publication. 11. Helsinki, Finland: Department of General Linguistics, University of Helsinki.

Koskenniemi, Kimmo. 1984. A General Computational Model for Word-form Recognition and Production: COLING '84, 178-181.

Koskenniemi, Kimmo, & K. W. Church. 1988. Complexity, Two-level Morphology and Finnish. Proceedings of the 12th International Conference on Computational Linguistitcs COLING-88, 335-340.

Lezius, Wolfgang, R. Rapp & M. Wettler. 1996. A morphology-system and part-of-speech tagger for German. In D. Gibbon, (ed.) Natural language processing and speech technology. Results of the 3rd KONVENS Conference, Bielefeld, October 1996. Mouton de Gruyter: Berlin.

Martí M. A. 1986. Un sistema de análisis morfológico por ordenador. Procesamiento del Lenguaje Natural, 4: 104-110

Meya, M. 1986. Análisis morfológico como ayuda a la recuperación de información. Procesamiento del Lenguaje Natural, 4:91-103.

Moreno-Sandoval, Antonio. 1991. Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español. Doctoral dissertation. Universidad Autónoma de Madrid: España.

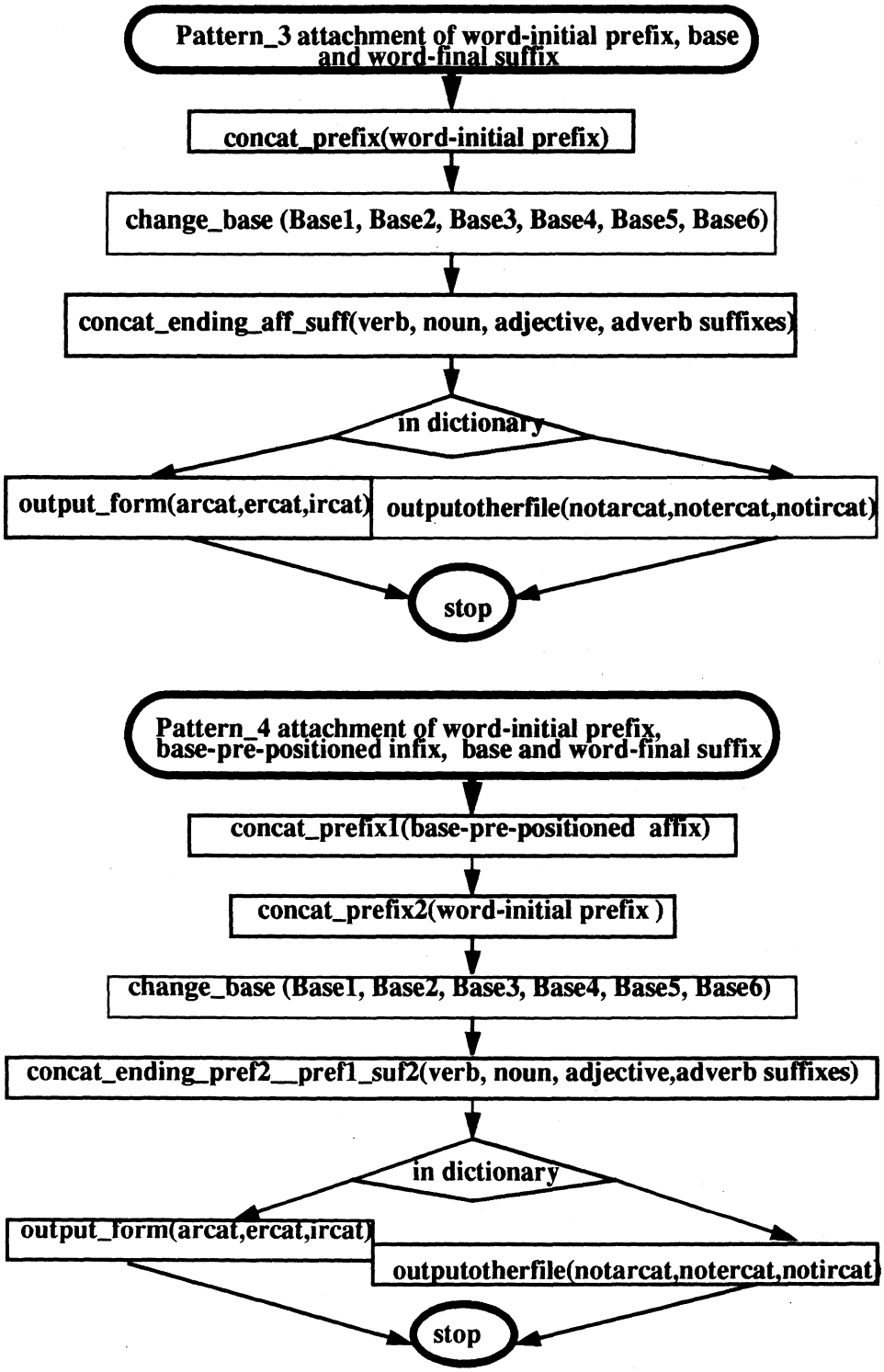
Nirenburg, Sergei et al. 1996. Mikrokosmos, Computing Research Laboratory, Las Cruces, NM: New Mexico State University

Ostler, Nicholas& B. T. S. Atkins. 1992. Predictable meaning shift: Some linguistic properties of lexical implication rules In J. Pustejovsky and S. Bergler (eds), Lexical Semantics and Knowledge Representation, 87-100. Berlin: Springer

Pollard, C., & I. Sag. 1987. An Information-based Approach to Syntax and Semantics: Volume 1 Fundamentals. CSLI Lecture Notes 13, Stanford CA. As cited in Viegas, Gonz'alez,

& Longwell, 1996.
Ritchie, Graeme R. 1992. Languages Generated by Two-Level Morphological Rules. Computational Linguistics. 18:41-59.
Ritchie, Graeme G. Russell, A. W. Black, & S. G. Pulman. 1992. Computational Morphology: Practical Mechanisms for the English Lexicon. ACL-MIT Press Series in Natural Language Processing. Cambridge, MA: MIT Press.
Russell, Graham, J. Carroll, & S. Warwick-Armstrong. 1991. Multiple default inheritance in a unification-based lexicon. ACL Proceedings, Berkeley, CA: ACL.
Sproat, Richard. 1992. Morphology and Computation. ACL-MIT Press Series in Natural Language Processing. Cambridge, MA: MIT Press.
Tzoukermann, Elizabeth, & M. Liberman. 1990. A finite-state morphological processor for Spanish. COLING-90. Helsinki, Finland, 3:277-282.
Viegas, Evelyne, M. C. González, & J. Longwell. 1996. Morpho-semantics and Constructive Derivational Morphology: A Transcategorial Approach. Technical Report MCCS-96-295. Las Cruces, NM: Computing Research Laboratory, New Mexico State University.
Viegas, Evelyne, B. Onyshkevych, V. Raskin and S. Nirenburg. 1996b. From Submit to Submitted via Submission: on Lexical Rules in Large-scale Lexicon Acquisition. Santa Cruz, CA: ACL-96 Proceedings.
Evelyne Viegas & S. Beale. 1996. Multilinguality and Reversibility in Computational Semantic Lexicons, INLG'96 Proceedings, Sussex, England: INGL.

Appendix A


Pattern_3 attachment of word-initial prefix, base and word-final suffix
concat_prefix(word-initial prefix)
change_base (Base1, Base2, Base3, Base4, Base5, Base6)
concat_ending_aff_suff(verb, noun, adjective, adverb suffixes)
in dictionary
output_form(arcat,ercat,ircat)
outputotherfile(notarcat,notercat,notircat)
stop
Pattern_4 attachment of word-initial prefix, base-pre-positioned infix, base and word-final suffix
concat_prefix1(base-pre-positioned affix)
concat_prefix2(word-initial prefix)
change_base (Base1, Base2, Base3, Base4, Base5, Base6)
concat_ending_pref2__pref1_suf2(verb, noun, adjective,adverb suffixes)
in dictionary
output_form(arcat,ercat,ircat)
outputotherfile(notarcat,notercat,notircat)
stop

Appendix B

GEN_WORD© The Tool: