Stephan Bopp
Institut für Informatik der Universität Basel

An Implementation of Italian Inflection and Word Formation

Abstract

This paper presents an implementation of Italian inflection and word formation with a system called Word Manager. The system is developed by a group of computer scientists, computer linguists and linguists at the Department of Computer Science, University of Basle, Switzerland. Word Manager is a system intended for the specification, use and maintenance of morphological dictionaries as described in Domenig/ten Hacken (1992). It has different subformalisms for the specification of morphological knowledge. All inflectional and word–formation processes described in the consulted grammars have been implemented by means of these subformalisms. At least as far as word formation is concerned, it is the at present most comprehensive specification of morphological knowledge of the Italian language.

1. Introduction: Word Manager

Word Manager is a system intended for the specification, use and maintenance of morphological dictionaries (see Domenig/ten Hacken 1992). It distinguishes two phases in the knowledge specification process: first the specification of lexeme classes and second the specification of instances of these lexeme classes. The knowledge about lexeme classes corresponds approximately to the morphological rule knowledge usually defined in the introductory section of dictionaries or in the morphology chapter in traditional grammars. It comprises, roughly speaking, the following definitions:

- how formatives are combined into wordforms and how wordforms are structured into inflectional paradigms defining a specific lexeme class (= rules for the generation of inflectional paradigms),
- how an individual instance of a lexeme class can be created (= knowledge about how entries may be added to a vocabulary),
- how formatives of lexeme classes and, optionally, word–formation affixes are combined into new formatives in order to create new instances of lexeme classes (= rules for word formation).

The specification of this first kind of morphological knowledge, the rule knowledge, is the object of the implementation of Italian morphology presented in this paper. The specification of instances of lexeme classes, the entry knowledge, will be carried out in a second step by a lexicographer.

Within the specification of the conceptual morphological knowledge, Word Manager distinguishes several subformalisms for the definition of different kinds of formatives and rules. *Inflectional Rules* define how *Inflectional Formatives* (e.g. stems, inflectional affixes) are combined into wordforms and wordforms into inflectional paradigms. *Word–Formation Rules* define how *Inflectional Formatives* and, optionally, *Word–Formation Formatives* (e.g. derivational affixes) are combined into new Inflectional Formatives. Furthermore, there are so–called *String Rules* which are similar —though not identical — to the two–level rules in Koskenniemi's two–level model (Koskenniemi 1983): they allow the formatives used for the construction of wordforms to be defined as linguistically motivated abstractions by ensuring that the combinations of formatives are mapped onto correct orthographic (surface) representations of wordforms.

The implementation of the Italian inflection and word formation comprises, under a strictly synchronical point of view, all the productive inflectional and word–formation processes of contemporary standard Italian described by the different sources (Garzanti 1987, Schwarze 1988, Serianni 1988).

2. Inflection

2.1. Nouns

Italian noun inflection is divided into different regular and irregular lexeme classes according to the inflectional suffixes the respective stems are combined with. Each lexeme class is defined by an Inflectional Rule (IRule) that specifies its inflectional paradigm. The nouns are only inflected for number. Generally, a gender feature is attributed individually to each entry. The IRule for the nouns of the +o/+i–class (e.g. 'ragazzo – ragazzi' *boy – boys*) combines a noun stem with the singular suffix '–o' to the singular form of the lexeme and the same stem with the plural suffix '–i' to its plural form. The citation–form is the wordform of the singular:

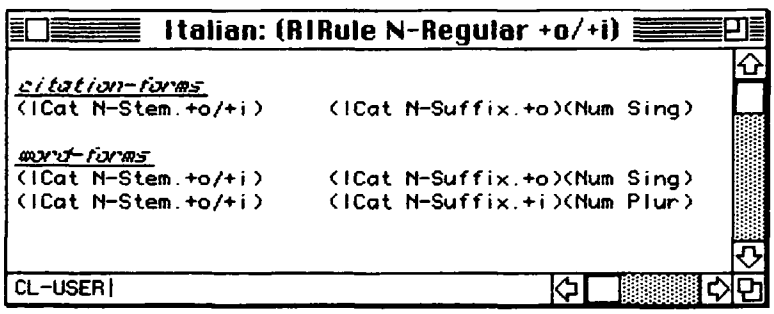

Fig. 1: Regular Inflectional Rule, Regular Nouns of the +o/+i–Class

The suffixes are defined as *fully specified formatives*, i.e. their strings and their qualifying features are defined within the specification of the rule knowledge:

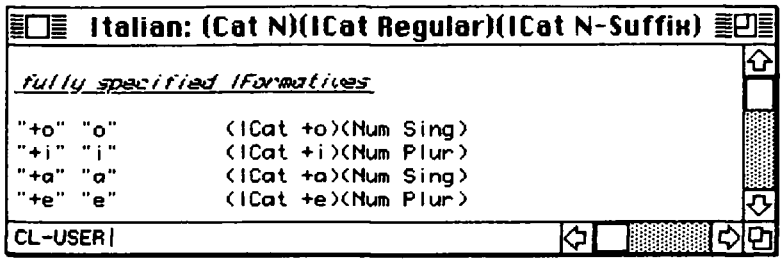

Fig. 2: Fully Specified Inflectional Formatives, Noun Suffixes

The stems are defined as *underspecified formatives*, i.e. only their qualifying features are defined within the specification of the rule knowledge:

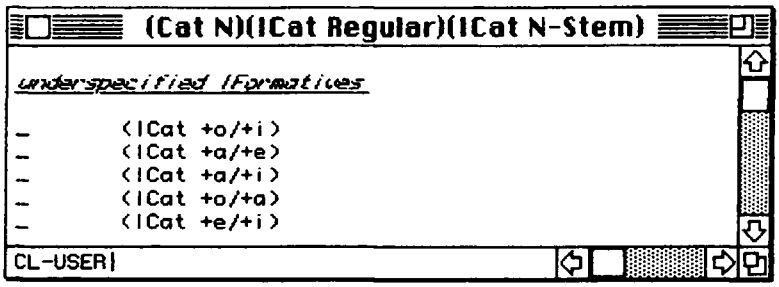

Fig. 3: Underspecified Inflectional Formatives, Noun Stems

The underspecified formatives are the key to the definition of how entries can be made to the dictionary. An IRule can only be fired when all formatives defined in it are fully specified. To make an entry into the above–mentioned noun class, one has to specify the strings of the underspecified formatives occurring in the IRule, i.e. the string of the noun stem. After the specification of the string"ragazz" for the stem (and a gender feature), the IRule has all the information it needs to generate the new entry:

"ragazzo" (Cat N)(Gender M)
"ragazz+o""ragazzo" (Num Sing)
"ragazz+i""ragazzi" (Num Plur)

The other regular noun lexeme classes are specified in a similar way. The most important ones are the +a/+e–class ('signora – signore'), the +e/+i–class ('studente – studenti') and the +a/+i–class ('problema – problemi'). Besides IRules for some minor lexeme classes like, e.g., invariable nouns ('il caffè – i caffè') or compounds inflecting inside the wordform ('cassaforte – casseforti' *safe – safes*), there are a few IRules for irregular nouns (e.g. 'uomo – uomini' *man – men*; 'bue – buoi' *ox – oxen*). These irregular nouns are specified as hard–coded entries.

Basically, the morphographic modifications connected with Italian noun inflection are handled by three String Rules and the introduction of three special lexical characters:

- Stems ending in '–c' /k/ or '–g' /g/ are rewritten '–ch' and '–gh' when combined with a plural suffix '–e' or '–i':
  'fuoc–i' –> 'fuochi' (plural of 'fuoco' *fire*)
  'drog–e' –> 'droghe' (plural of 'droga' *drug*)

With some stems, however, the insertion of the diacritic 'h' does not occur. They are specified with the special lexical characters 'C' and 'G', which do not match the conditions formulated in the String Rule for the 'h'–insertion. By another String Rule they are rewritten 'c' and 'g' when mapped onto the surface representation:

'amiC–i' –> 'amici' (plural of 'amico' *friend*).

- Stems ending in unstressed or diacritic 'i' are rewritten without the 'i' when combined with '–i':
  'studi–i' –> 'studi' (plural of 'studio' *study*)
  'baci–i' –> 'baci' (plural of 'bacio' *kiss*)

Stems ending in a stressed 'i' (not marked in orthography) keep it before '–i'. They are specified with the special character 'í' that does not match the conditions formulated in the first String Rule. By a second String Rule 'í' is rewritten 'i' when mapped onto the surface representation:

'pendí–i' –> 'pendii' (plural of 'pendio' *slope*)

- Stems of the +a/+e–class ending in '–ci' or '–gi' (diacritic 'i') preceded by a consonant are rewritten without the diacritic when combined with '–e':

  'spiaggi–e' –> 'spiagge' (plural of 'spiaggia' *beach*)
  'camicie' –> 'camicie' (plural of 'camicia' *shirt*).

Stems of the same orthographic shape but with a stressed 'i' are again specified with the lexical character 'í':

'liturgí–e' –> 'liturgie' (plural of 'liturgia' *liturgy*).

For the specification of the Italian noun inflection, 21 IRules (6 of which for hard–coded entries) and 3 String Rules were needed.

2.2. Other parts of speech

The lexeme classes of all other parts of speech are formalized analogously. IRules combine stems with the respective inflectional suffixes. Morphologically irregular lexemes are specified as hard–coded entries. Morphographic modifications are specified by String Rules. The lexeme classes of non–inflecting parts of speech like, e.g., prepositions and conjunctions are defined by IRules that specify an inflectional paradigm containing one single word form consisting of one single formative.

3. Word formation

Word–Formation Rules (WFRules) define which Inflectional Formatives (IFormatives) out of which IRules can be combined with each other or with Word–Formation Formatives (WFFormatives) to new IFormatives of new lexemes.

3.1 Derivation

Derivation is formalized by WFRules that combine IFormatives (usually stems) with WFFormatives (derivational affixes) to new stems belonging to new lexeme entries. All WFFormatives are defined within the specification of the rule knowledge as fully specified formatives. An example of a WFRule defining derivational processes is the following rule for the noun–to–noun derivation by suffixing:

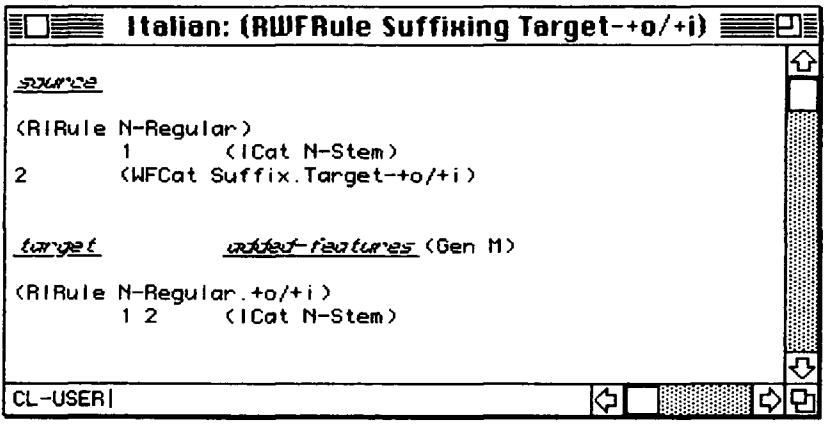

Fig. 4: Regular Word–Formation Rule, Noun–to–Noun Suffixing

A noun stem (ICat N–Stem) belonging to a lexeme class defined by any IRule for regular nouns (RIRule N–Regular) and a derivational suffix qualified by the feature set (ICat N–Suffix.Target–+o/+i) are combined in the order defined by the digits representing them under *target*. The combination is qualified as a noun stem belonging to the +o/+i–class. The target IRule, i.e. the IRule in which the new formative is entered, is the (RIRule N–Regular.+o/+i) described in Section 2.1.

When a new entry to the dictionary is made by the application of this WFRule, the stem of a regular noun (e.g. 'bibliotec' of 'biblioteca' *library*) and one of the the suffixes qualified by the requested feature set (e.g. '–ari') are combined to a new noun stem ('bibliotecari') that is entered into the target IRule. Since the noun stem is the only underspecified formative in this IRule, the specification of its string is the information needed for the IRule to be fired:

"bibliotecario" (Cat N)(Gen M)	*librarian*
"bibliotec–ari–o""bibliotecario"	(Num Sing)
"bibliotec–ari–i""bibliotecari"	(Num Plur)

This illustrates that there are two ways of entering a new lexeme to the dictionary: 1) simple entries by direct *firing* of an IRule and 2) complex entries by indirect *firing* of an IRule through the firing of an WFRule.

The morphographic modifications occurring with suffixing are specified by String Rules. E.g.

- All noun, adjective and verb stems ending in '–c' /k/ or '–g' /g/ are rewritten '–ch' and '–gh' when combined with a suffix beginning with 'e–' or 'i–':
  'bosc–o' >'bosch–iv–o' (*wood, forest* > *wooded, wood...*)
  'pieg–are' >'piegh–evol–e'(*fold* > *foldable*)
- All noun, adjective and verb stems ending in '–c' /tʃ/ or '–g' /–dz, / are rewritten '–ci' or '–gi' when combined with a suffix beginning with 'a–', 'o–' or 'u–':
  'voc–e' >'voci–on–e' (*voice* > *deep, strong voice*)
  'dolc–e' >'dolci–astr–o'(*sweet* > *sweetish*)
- All noun, adjective and verb stems ending in a diacritic or unstressed '–i' are rewritten without the final 'i' when combined with a suffix beginning with 'i–':
  'doppi–o' >'dopp–ist–a' (*double* > *player in a doubles pair*)
  'lasci–are' –> 'lasc–it–o' (*let* > *legacy*)

All derivational processes occurring in modern standard Italian have been formalized in a similar way. The derivation part of the specification is subdivided in ten units according to the parts of speech of the source entry and its derivative (N>N, N>A, A>N, A>A, V>N etc.). Within these units,

rules are formulated for the different relevant types of processes like suffixing, prefixing and conversion (or zero derivation) as well as for the possible combinations between these types of processes (e.g. 'bianc–o > im–bianch–ire' *white* > *whiten*). The specification of the individual rules is further affected by the inflectional class of the source entries and the derivatives, so that a total of 114 WFRules and 73 String Rules had to be formulated.

3.2. Compounding

The Italian compounding is quite complex, too. The complexity is due to various interacting facts:

- Generally, not stems but full wordforms (including the inflectional suffixes) are combined.
- Compounds can be inflected by the modification of the inflectional suffix of the first element, the second element, both elements or none of the elements, depending on the relation between the two source lexemes (head and modifier, endocentric vs exocentric compounds, etc.).
- Depending on the degree of semantic fusion of the two elements (lexicalization of the compound), several kinds of formal fusion phenomena occur quite easily and can therefore be considered productive: e.g. the suffix of the first element is invariable even though the character of the compound requires it to be variable ('mezzogiorno' – 'mezzogiorni' *midday*).

In the specification of WFRules for compounding, a first distinction has to be made between 1) compounds that inflect inside the wordform, 2) compounds that inflect by modifying only the suffix of the second element and 3) invariable compounds. For the first group special IRules are formulated in which the new formatives are entered. They define an inflectional paradigm characterized by modifications inside the wordforms ('capostazione – capistazione' *stationmaster*). The compounds of the second group can be inserted into the regular IRules because, modifying only the suffix at the right end of the wordform, they inflect like simple entries ('mezzogiorn–o – mezzogiorn–i' *midday*). The invariable compounds are entered into IRules for invariable lexemes ('il/ i portachiavi' *key ring(s)*).

Within these groups different types of WFRules have to be formulated according to the parts of speech the compounds are composed of (N+N, N+A, V+N, etc.). Since, generally, whole wordforms are combined, the specification of the source formatives has to include not only stems but suffixes as well. Therefore, the subgroups are further subdivided in WFRules defining the relevant possible combinations of inflectional class, gender (e.g. for the gender agreement between adjectives and nouns: 'mezzogiorno'

midday vs 'mezzanotte' *midnight*) and number (e.g. for invariable compounds: 'coprifiamma' *flash hider* vs 'lanciafiamme' *flame-thrower*).

Proceeding in this way, all the productive compounding processes we could trace were specified by 37 WFRules (N+N: 4, N+A: 10, A+N: 9, A+A: 4, V+N: 10) and 6 String Rules.

4. Conclusions

The Word Manager formalism allowed to formalize all the productive inflectional and word-formation processes of contemporary standard Italian. Where the information in the consulted literature was incomplete or lacking (especially concerning morphographic phenomena connected with word formation) we tried to find regularities by examining a large number of examples. In this way, the formalism was not only used to formalize the morphological knowledge described in the grammars, but also to gain some new information on Italian morphology. It proved to be powerful enough to cover the morphology — including the word formation — of a language with a great variety of forms like Italian. A total of 383 different rules (IRules, WFRules and String Rules) were formulated and implemented. Therefore, our specification is the at present most comprehensive implementation of morphological knowledge of the Italian language.

After the formalization of the Italian morphology, an equally comprehensive formalization of the morphology of German has been completed. The results are just as positive: the formalism proved to be powerful enough to cover the morphology of a Germanic language, including the complex phenomena connected with German composition. Smaller implementations of French, English and Dutch morphologies showed equally satisfying results (e.g. Brunner 1991, Garcia 1991, Gregorio 1993, Gupta 1989).

5. Prospects

The Italian and German databases are being tested and prepared for the construction of large morphological dictionaries. A new tool — called *Phrase Manager* — has been added to the Word Manager formalism. It allows the formalization, registration and the recognition of multi-word-units. In our definition, multi-word-units are — roughly speaking — lexical units consisting of more than one orthographic word. On the basis of the information stored in a Word Manager database, all kinds of multi-word-units, from analytical inflectional forms like 'have been', 'werde gehen', etc. to idiomatic expressions like the well-known 'to kick the bucket', can be captured with all their morphosyntactic modifications. The prototype version of Phrase Manager has recently been completed (Pedrazzini 1994) and is currently being tested on all kinds of German and Italian idiomatic expressions.

References

Bopp, S. 1993. *Computerimplementation der italienischen Flexions– und Wortbil–dungsmorphologie.* Hildesheim: Olms Verlag.

Brunner,C.1991. *An Implementation of English Morphology Using the Program Word Manager.* Lizentiatsarbeit an der Philosophischen Fakultät I der Universität Zürich.

Domenig, M. 1990. "Lexeme–Based Morphology: A Computationally Expensive Approach Intended for a Server–Architecture". *Proceedings of the 13th International Conference on Computational Linguistics, COLING–90.* Helsinki, August 20–24.

Domenig, M. and ten Hacken, P. 1992. *Word Manager: A System for Morphological Dictionaries.* Hildesheim: Olms Verlag.

Garcia,C.1991. *Computerimplementation der deutschen Morphologie,* Lizentiatsarbeit an der Philosophischen Fakultät I der Universität Zürich.

Garzanti 1987. *Il grande dizionario Garzanti della lingua italiana.* Milan: Garzanti Editore.

Gregorio,S.1993. *Implementation of English Inflectional and Derivational Morphology.* Lizentiatsarbeit am Institut für Informatik der Universität Basel.

Gupta,A.1989. *La formalisation de la morphologie française sur la base du système Word Manager.* Lizentiatsarbeit an der Philosophischen Fakultät I der Universität Zürich.

Koskenniemi, K. 1983. *Two–Level–Morphology: A General Computational Model for Word–Form Recognition and Production.* Doctoral thesis, University of Helsinki, Publications Nº 11.

Pedrazzini, S. 1994. *Phrase Manager, A System for Phrasal and Idiomatic Dicitionaries.* Hildesheim: Olms Verlag.

Schwarze, C. 1988. *Grammatik der italienischen Sprache.* Tübingen: Niemeyer.

Serianni, L. 1988. *Grammatica italiana.* Torino: UTET.