

ANNOTATING LARGE CORPORA FOR STUDYING ITALIAN DERIVATIONAL MORPHOLOGY

NICOLA GRANDI

FABIO MONTERMINI

FABIO TAMBURINI

ABSTRACT: This paper presents ongoing research concerning the annotation of large corpora with morphological information. It aims at providing a general schema for inserting rich morphological information to enable complex corpus queries of word internal structure. Annotating real corpus data presents challenges that can hardly be managed with traditional linear analysis of word structure, but can efficiently and correctly be handled with different, more complex, structures. For this reason, we propose Derivation Graphs as a new tool for representing the structure of complex words, and we discuss the theoretical consequences of this choice on the representation of affixes, a crucial issue for all morphological models.

KEYWORDS: corpus annotation, derivational morphology, derivation graphs, representation of affixes.

1. THE REPRESENTATION OF AFFIXES*

The nature of affixes and, more generally, of bound morphological items constitutes a problematic issue for all models of morphology. Similarly, the representation of these elements in dictionaries and other reference books is not a trivial issue, and crucially depends on the theoretical perspective adopted.

In the last few decades, most morphological theories treated affixes as (non autonomous) lexical objects. Since Aronoff's (1976) seminal work, affixes have been represented as part of "Word Formation Rules" (henceforth WFR) with a linear arrangement and a set of constraints and conditions on the input and the output:

- (1) $X]_N -aI]_A$
Condition: $X \neq [Y]_{V_{ment}}$ (Aronoff, 1976: 54)

* We are grateful to Malvina Nissim for having discussed with us some of the issues presented in this paper, and to two anonymous referees for their comments on a previous version of the paper.

- (2) $[[]_N + oso]_A$
 [+com]
 [-anim] (Scalise, 1994: 99)¹

Constraints and conditions delimit the number of items that can replace the variable indicated by X (or by an empty space in Scalise's representation). So, the English suffix *-al* selects nominal bases to form adjectives (with the exception of deverbal nouns in *-ment*, i.e. bases with the structure $X_{\sqrt{ment}}$: *ornamental* vs. **employmental*). The Italian suffix *-oso* forms adjectives from common and inanimate nouns. In the background, a set of general conditions on WFRs, such as the Unitary Base Hypothesis, the Unitary Output Hypothesis, Blocking, etc., is at work.

This is not the place to present an exhaustive and critical survey of all the possible consequences of this representation and of the nature of the items that can replace the variables (are they words, lexemes, roots, stems, etc.?). In a trivial and oversimplified way, we can view the idea underlying most of these theories as a correspondence:

- (3) one input → one morphological operation → one output → one meaning

The advent of new technologies, the development of techniques of computational processing, and the interaction with corpus linguistics introduce new challenges to morphology. Theoretical models can be tested against the evidence coming from the analysis of large amounts of corpus data, which cast new light also on the issue of the representation of affixes and WFRs. Unlike lists of words coming from dictionaries or lexicographical inventories, corpora provide “real” data from authentic linguistic contexts. In this scenario, representations such as those in (1) and (2) naturally became the keys for searching corpora in order to extract the occurrences of derived words. In other words, the components of those representations (input and output category, affix position, etc.) were used to set search parameters.

The possibility of automatically extracting large amount of data from corpora and observing them in their real contexts rather than in isolation has created many difficulties for the maintenance of this theoretical framework, both in a practical and theoretical perspective:

- i) The bijective mapping between a morphological operation and an output does not hold: there are words that behave differently, according to the context of occurrence:²

¹ Cf. also Matthews (1974: 63): $[X]_V \rightarrow [X + \text{'[jən]}]_N$.

² Italian data are from CORIS/CODIS, a 130-million-word corpus representative of written contemporary Italian (Rossini Favretti, Tamburini & De Santis, 2002).

- (4) a. *Suo padre, modesto calciatore in gioventù, è operaio alla società dei telefoni, alza spesso il gomito*
 ‘His father, an average football player in his youth, is a factory worker in the telephone company, he tips the bottle’
- b. *Il bagno della vecchia casa di periferia in un quartiere operaio di Praga era meno ipocrita*
 ‘The bathroom of the old suburban house in a working class area of Prague was less hypocritical’

In (4a) *operaio* ‘factory worker’ is, as expected, a typical noun. In (4b) it modifies the noun *quartiere* ‘district’: a *quartiere operaio* designates a working-class area. When it is used as a modifier of another noun, *operaio* behaves like an adjective, and can agree, for instance, with the head noun in gender and number:

- (5) *Da giovane si fece le ossa partecipando alle lotte operaie di Torino, quelle dei primi anni 60.*
 ‘As a young man he gained experience participating in Turin’s workers’ struggles, the ones from the early Sixties’

Therefore, it is often the context of occurrence that ultimately allows us to tag a word for its part of speech.

ii) The bijective mapping between a morphological operation and an input does not hold either: in some derived words the category of the base can hardly be identified (6a); in other, well-known cases the same affix can select bases of different categories, with the same output (6b):³

- (6) a. *vecchiaia* ‘old age’: [[vecchio]_N + aia]_N or [[vecchio]_A + aia]_N
fischietto ‘whistle’: [[fischio]_N + etto]_N or [[fischiare]_V + etto]_N
- b. *mangiabile* ‘eatable/edible’: [[mangiare]_V + bile]_A
 vs.
papabile ‘likely to become a pope’: [[papa]_N + bile]_A

iii) A single derived word can show different derivational paths, if two or more affixes are present (the concurrent presence of prefixes and suffixes leads to such phenomenon). Let us consider the case of *scomponibile* ‘decomposable’. This form can be analysed as containing the stem *componi* (corresponding to the verb *comporre* ‘to compose’), a reversative prefix *s-* and an adjectival suffix *-bile*. The two intermediate forms, the verb *scomporre*

³ In the representation of WFRs, bases are put in square brackets and are indicated with their citation form, that does not necessarily correspond to the actual stem on which the rule operates. This is particularly evident for verbal bases, for which the citation form corresponds to the infinitive, a form that is never, in itself, the base of a WFR.

‘decompose, dismantle’ and the adjective *componibile* ‘composable’, are both attested words. It is impossible to determine, on empirical grounds, whether the adjective *scomponibile* is derived from the adjective *componibile* by adding the prefix *s-* or from the verb *scomporre* by adding the suffix *-bile*. Since linearity does not allow to prefer one analysis over the other, the choice can only be made on a theoretical basis. However, the very fact that there must be an unequivocal answer is already a theoretical choice.

The vast majority of the data presented so far was usually considered as “peripheral” or “non prototypical” in many morphological theories. However, corpus based analyses reveal that they are less “odd” (or at least less rare) than one would think. Situations as those mentioned in (i)-(iii) cannot be directly accounted for by the representations in (1) and (2). In this picture, we observe that one dimension (i.e., the necessity of arranging morphological items in a linear order) is not sufficient to represent the complexity of word formation processes. Moreover, not all the information can be projected on the abstract level. There is an amount of information that is not constant and invariable, but strictly depends on actual use (and would therefore be better represented by variables), and this amount is probably higher than expected.

What clearly emerges is that the way in which all these situations are represented within a theoretical framework not only addresses the solutions of these problems, but necessarily conditions and pre-determines the choices on an operational level. In other words, corpus annotation cannot be considered as being impervious to the theoretical framework the researchers adopt. Conversely, annotating a corpus that is suitable for morphological investigations implies choices that have inevitable theoretical consequences. The design of a corpus annotation, and in particular of its search options, is directly conditioned by such problems as those listed in (i)-(iii) above. In particular, one could be interested in finding out all the possible “nuances” of an affix and of a word formation process and not to exclude *a priori* a (sub)set of possible occurrences. This raises crucial theoretical issues concerning, for instance, the very nature of affixes and the way of representing them, the ratio between the information which is constant and “context-proof” and the information which corresponds to variables whose actual values can only be determined in their context of occurrence (see also Celata & Bertinetto, 2010 for a discussion about these problems for Italian). More generally, it is commonly assumed in corpus linguistics that any corpus annotation, whichever linguistic level is chosen, is the projection of a specific theory and cannot be trusted as an instance of “real/authentic uncontaminated data”.

This paper presents some general considerations about affix representation, both in an operational and a theoretical perspective,

which are the outcome of a process of “restyling” of the corpus CORIS/CODIS (cf. footnote 2) carried out by the authors and aimed at adapting it to fine morphological investigations. The observations we propose are the outcome of our efforts at solving some practical problems related to the morphological analysis of words in the corpus, that include bringing out their internal structure, making the information associated to all the constituents of a complex word (and not just those in its external node) accessible, and highlighting all the possible derivational paths of a single complex word. In the next section, we present the solution we have adopted for some of these operational problems, and we suggest a new possible representation of word formation processes that goes beyond the linearity constraint: the “Derivation Graph”. The last part of this paper is devoted to a broader discussion on the theoretical consequences of some of these choices within the more general picture of current research in morphology.

2. MORPHOLOGICAL ANNOTATION

Morphological analysis performed with computational tools is a fundamental step in various natural language processing (NLP) tasks, for example: a) Part-of-Speech tagging (see the EVALITA 2007 and 2009 evaluation campaign for Italian⁴); b) Parsing & Grammar checking; c) Spell checking & correction; d) Information retrieval (stemming & lemmatisation); e) Text-to-speech (prosody generation), etc.

Unlike other NLP tasks, largely dominated by machine-learning techniques, computational morphology relies mainly on rule-based systems that apply deterministic (opposed to statistical/stochastic) methodologies.

A review of these systems and of the associated techniques is outside the scope of this paper. It will suffice to say that, after the introduction of the two-level morphology by Koskenniemi (1983), the reference formal model for this kind of systems are finite state transducers (FST). These models implement two different operations: a) analysis, which extracts all the information connected with a word form associating it to a standardised notation “lemma+features” – for example the form *libri* (‘books’) becomes “*libro*+N+Masc+Plur” and the form *amo* (which is ambiguous in Italian, and may correspond to ‘I love’ or to ‘hook’) is associated to two different lemmas, “*amare*+V+Ind+Pres+1ps” and “*amo*+N+Masc+Sing” – and b) generation, the opposite operation, which associates to a structure “lemma+features” the corresponding word form – for example the structure “*dormire*+Verb+Ind+Pres+1ps” is associated to the word form *dormo* ‘I sleep’.

⁴ <http://www.evalita.it/>.

Italian is a language with a complex inflectional system and with a rich derivational morphology. The pool of computational resources devoted to this language is overall not very rich. In particular, there do not seem to exist freely available morphological analysers covering a large part of the Italian lexicon. There are some specific tools that handle morphological information by implementing the described morphological operations – Morph-It (Zanchetta & Baroni, 2005), MAGIC (Battista & Pirrelli, 2000), TextPro/MorphoPro (Pianta, Girardi & Zanoli, 2008) – and complete parsing systems able to handle morphological information – TUT parser (Lesmo & Lombardo, 2002), GETARUN (Delmonte, 2000) – but none of them embodies a large lexicon (more than 100.000 lemmas) and is freely available.

Scholars at the Dipartimento di Studi Linguistici e Orientali (DSLO), University of Bologna, are currently developing a large morphological analyser – AnIta – based on a very large lexicon (about 120.000 lemmas): this tool relies on a powerful package – HFST (Lindén Silfverberg & Pirinen, 2009) – designed for the implementation of morphological analysers and other tools which are based on weighted and unweighted finite-state transducer technology and can be easily extended to include various types of specific morphological information.

In order to provide operating support for morphological research, we need to annotate CORIS/CODIS with a rich set of morphological information allowing the user to cast complex queries to the corpus in order to retrieve all the data combinations needed for his/her studies. A powerful and rich morphological analyser, such as AnIta, is a fundamental tool for a successful annotation of large corpora with morphological information.

In the next sections we describe two different morphological annotation schemas suitable for marking different levels of morphological processes that we are introducing, first in AnIta, then as annotation streams inside CORIS/CODIS.

2.1 The first approximation: form segmentation

We devised a first level of annotation able to mark the internal segmentation of word forms. Each form will be associated with a linear structure that can be described by the following regular expression schema:

$$/(PREF>)*BASE(<SUFF)*(-INFLEND)?/$$

where PREF, BASE, SUFF and INFLEND are strings that represent a prefix, a base, a suffix and an inflectional ending, respectively. The following examples can describe this process:

- (7) *dis>continu-i* 'discontinuous' (ADJ, MASC, PL)
an>nota<zion-e 'annotation' (N, FEM, SING)
ri>ab>bassa<ment-o 'relowering' (N, MASC, SING)
in>arid-irono 'they dried up' (V, REMOTE PAST, 3rd PL)

The insertion of this annotation allows for a number of sophisticated queries to corpus data by using regular expressions, for example:

- (8) */^dis>./* word forms prefixed with *dis-*
/.+<on-[eia]/ word forms suffixed with *-one* (*-oni*, *-ona*)
/^in>.+<ità/ word forms simultaneously prefixed with *in-* and suffixed with *-ità*, such as:
inabitabilità 'unsuitability to being occupied'
incongruità 'incongruity'
inospitalità 'inhospitality'
 ...

Since the actual lexicon of Italian is a complex and historically stratified entity, the exact segmentation of each single word is not a trivial task. However, in order to determine whether a complex word should be segmented or not we followed some basic – and possibly uncontroversial – principles: a) a complex word is segmented only when the base is a clearly recognisable autonomous Italian word; thus, all the forms derived from Latin, Ancient Greek or other languages that are only semi-transparent (e.g. *im-berbe* 'beardless', *catech-ismo* 'catechism', *de-clinare* 'decline, refuse')⁵ are not segmented; b) affixes are kept without modification; any formal variation (vowel or consonant erasure, gemination, epenthesis, etc.) is ascribed to the base, as the following examples show:⁶

- (9) *anti+incendio: anti>ncendio* (PREF+fire: 'fireproof')
contro+ordine: contro>rdine (PREF+order: 'counter-order')
gloria+oso: glori<oso (glory+SUF: 'glorious')
trans+sessuale: trans>essuale (PREF+sexual: 'transexual')
ap+prendere: a>pprendere (PREF+take: 'to learn')
città+d+ino: cittad<ino (city+EPENTH+SUF: 'urban / citizen')

While this first level of morphological annotation allows for a large number of complex queries, it is still unsuitable to represent some fundamental information. First of all, it does not contain any indication

⁵ In these forms the boundary between the base and the affix is neatly recognizable. The affix is still productive, but the base does not correspond to an actual word of Italian.

⁶ In these examples, and in the following ones, we omit marking inflectional endings, a detail which is not relevant for the present discussion.

about the lexical class of the bases and of the derived forms and, secondly, the representation of Italian complex words it provides is not detailed and powerful enough. A more complete annotation schema, able to complete this first level segmentation, has to be devised in order to capture the complex details of morphological processes.

2.2 *The proposed solution: Derivation Graphs*

As stated above, two problems are pressing while annotating a corpus. First of all, the derivational processes underlying some word forms cannot be easily described as single derivational trees; instead, a single derived word can involve different possible interpretations giving rise to different trees; consequently, a one dimension model is unsuitable to account for such complex words. Moreover, in order to be able to retrieve all possible morphological combinations, we need to incorporate into the corpus annotation information about the lexical classes both of the bases and of the complex words derived by affixation and to make it available for the users.

We will present the proposed solution to these problems by discussing some examples. Let us consider again the complex word $s > compons < bile$ ‘decomposable’, already mentioned in section 1. As said above, this form can be described as the result of two possible derivational paths, and, consequently, it can be represented by two different trees (we represent the tree using the parenthesised notation indicating the class of the derived form as a subscript):

$$(10) \quad \begin{array}{l} [[s > compons_V]_V < bile]_A \\ [s > [compons_V < bile]_A]_A \end{array}$$

Choosing one of these options, and, consequently, discarding the other, is a strong theoretical choice, since, as we asserted in iii) above (section 1), it is impossible to determine, on empirical grounds, whether the adjective *scomponibile* is derived from the adjective *componibile* by adding the prefix *s-* or from the verb *scomporre* by adding the suffix *-bile*. We are firmly convinced that, in this case (and in all similar cases), a corpus should not impose just one view, but should make all the possibilities available to users.⁷ Therefore, the crucial point is how to compact the two interpretations in (10) into one single formal structure able to correctly encode all this information.

The formal structure that naturally extends a tree structure is the “graph”. If we consider each element intervening in a derivational process (the base and the affix(es)) as the nodes of a graph (keeping the information

⁷ Of course, also this choice has strong theoretical consequences; we will discuss them in the next sections.

on the nature of the affix, as in the segmentation annotation) and the “derivation relation” as the formal device for defining the edges of the graph, we can build the “Derivation Graph” (DG) for the form *scomponibile* as in Figure 1c. The edges have arrows which mark the direction in which a derivation can take place, starting from the affix and ending at the base to which it directly applies.⁸ As for complex words showing the simultaneous presence of two or more affixes, the DG represents all the possible derivational paths; i.e., it does not imply that the affixes are attached to the word only in one predetermined order. Note that the class of the derived word is indicated on the edge, to emphasise the fact that it is the process that produces a specific lexical class and not the affix itself (affixes, especially prefixes, can form words of different classes).

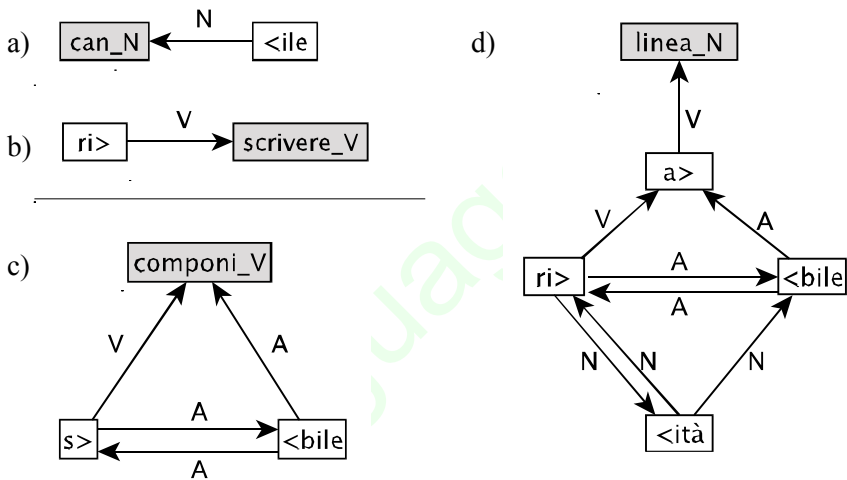


FIGURE 1. DERIVATION GRAPHS FOR THE FORMS *CANILE* ‘DOG POUND’, *RISCRIVERE* ‘TO REWRITE’, *SCOMPONIBILE* ‘DECOMPOSABLE’ AND *RIALLINEABILITÀ* ‘POSSIBILITY OF BEING REALIGNED’. THE SHADED NODE REPRESENTS THE BASE.

In order to navigate a graph, two rules must be obeyed:

- a) the starting point is always the base, that is the upper element (highlighted with grey in Figure 1);⁹
- b) each edge must be always travelled in the opposite direction of the arrow.

⁸ This statement suggests that the derivational process has a specific direction: from the affix to its base. In other words, that an affix selects the base. Of course this statement would deserve much more attention, but it is outside the aim of this paper and it will be faced in a further step of this research.

⁹ So, in complex graphs like the one in 1c, the prefix cannot be associated with the suffix (or viceversa) before being associated with the base.

Therefore it is possible to reconstruct all the possible interpretations of a derivational process by navigating the DG following a simple rule: every path in the graph starting from the base and built reversing the derivation relation (i.e. traveling the edges in the opposite direction of the arrows) that includes all the nodes at once leads to a possible interpretation of the derivational history of a complex word, and produces a tree describing this process. Looking at Figure 1c, it is very simple to recognise only two different paths, each corresponding to one of the trees in (10). For example, starting from the base, and travelling the left edges (in the opposite direction of the arrow), we assume that *componi* associates to *s-*, forming the prefixed verb *scomporre*; then, travelling the lower edge, we link this verb to the suffix *-bile*, obtaining the adjective *scomponibile*. This path includes all the three nodes of the graph at once, and corresponds to the pattern in (10a).

Let us consider now a more complex example, the form *ri>al>linea<bil<ità* ‘possibility of being realigned’. In this case, there are three possible analyses for the derivational process, each one corresponding to a different tree:

- (11) a. $[[[ri>[al>linea_N]_V]_V<bil]_A<ità]_N$
- b. $[ri>[[[al>linea_N]_V<bil]_A<ità]_N]_N$
- c. $[[ri>[[al>linea_N]_V<bil]_A]_A<ità]_N$

Navigating the DG in Figure 1d yields all and only the three possibilities listed in (11). Figure 2 shows with thick lines the path in the graph leading to the interpretation (11b).

Note that in the DG we chose to normalise all the affixes to their standard form, removing any variation and indicating the inflectional ending of the quotation form.

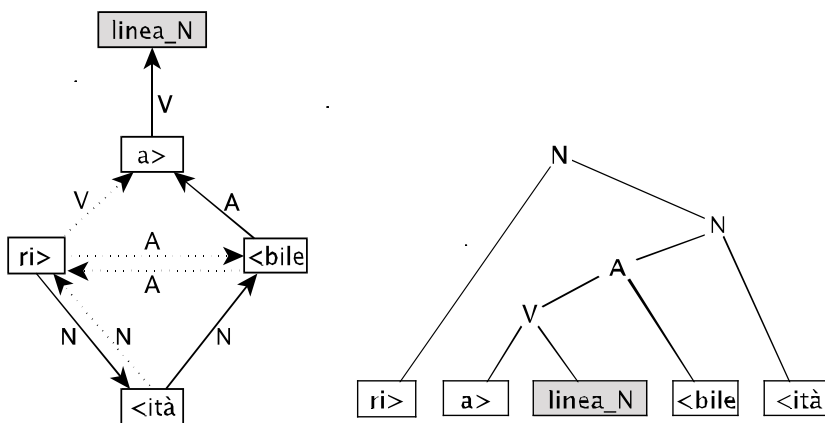


FIGURE 2. DG FOR THE WORD *RIALLINEABILITÀ* ‘POSSIBILITY OF BEING REALIGNED’ OUTLINING THE INTERPRETATION IN (11b) AND THE RESULTING TREE.

The DG we propose is a way of formalizing the totality of the derivational possibilities connected with a complex word. In that sense it can be considered a theoretical extension of the derivational trees traditionally used to express such linguistic phenomena.

One of the advantages of proposing such DGs is that they can be used fruitfully in corpus annotation as a base for designing morphologically complex queries on corpus data. To do this, it is necessary to incorporate the information linked with the DG into a simple but effective annotation schema allowing complex queries to be performed such as:

- retrieving all the occurrences of the suffix X where a noun is created from an adjective;
- retrieving all the prefixes that take a verb as base;
- listing all the affixes that form nouns from a specific base;
- ...

From a theoretical/computational point of view there are various ways of representing a graph structure, depending on the intended final use of such information. One of these methods consists in listing all the graph edges. Using this representation we can describe an entire graph as a single string considering the concatenation of the following two edge schemas:

$$\text{prefix}_{DC} > \text{base}_C \qquad \text{base}_C < \text{suffix}_{DC}$$

where *DC* is the derived lexical class (which has been moved from the edge to the affix governing the derivation, in order to keep the string readable and simple), and *C* is the class of the base. For example, the DG in Figure 1c can be expressed through the following list of its edges:

$$s_V > \text{componi}_V \mid \text{componi}_V < \text{bile}_A \mid s_A > \text{bile}_A \mid s_A < \text{bile}_A$$

where the character ‘|’ acts as a separator between the edges.

Once each word form in the corpus has been annotated with the string expressing the DG associated with it, the construction of simple but extremely powerful queries is possible with any corpus management program permitting the use of regular expressions in corpus queries, such as the IMS/Corpus Workbench. Some examples of these queries are given below:

- $/\cdot +_V < \text{bile}_A/$ all the instances/concordances in which the suffix ‘<*bile*’ forms an adjective from a verb;
- $/s_A > \cdot +_A/$ all the instances/concordances in which the prefix ‘*s*’ forms an adjective from another adjective;
- $/\cdot +_V < \cdot +_A/$ all the instances/concordances in which a suffix forms an adjective from a verb.

In Italian, it is common that a single affix derives words belonging to different lexical classes. It is the case, for example, of the word

$[oper_N<ai0]_{N/A}$ mentioned in (4) (section 1). In order to take these cases into account, we propose to encode all the possible combinations of the four major lexical classes (N, A, V, D (=adverb)) by using the simple encoding schema depicted in Table 1. So, a problematic word like *operaio* can be associated with the structure $[oper_N<ai0]_C$. In this way, *operaio* will be included in the results from queries aimed at extracting derived nouns and derived adjectives from the corpus.

CODE COMBINAT.	CODE COMBINAT.	CODE COMBINAT.	CODE COMBINAT.
A A	E A+V	I A+D+N+V	N N
B A+D	F A+D+N	J D+N	O N+V
C A+N	G A+D+V	K D+V	V V
D D (Adv)	H A+N+V	L D+N+V	

TABLE 1. ENCODING SCHEMA FOR EXPRESSING ALL THE POSSIBLE COMBINATIONS OF MULTIPLE WORD-CLASSES.

The class encoding schema we propose covers all the combinations that are logically possible, although most of them are not attested in Italian.¹⁰ Again, a system based on regular expression searches will help find all the possible combinations while querying the corpus.

3. THEORETICAL CONSEQUENCES

As can be seen from the previous sections, the interaction with computational linguistics, and in particular with corpus studies, has two important consequences for morphological theory. On the one side, new theoretical questions and proposals may emerge from the need to implement models of large-scale corpus analysis. On the other side, new technologies allow for quick access to a quantity of annotated data that was inconceivable some decades ago. Collecting large amounts of lexical data is essential for the study of morphological processes, and specifically of derivation. In particular, corpora make the access to neologisms and rare derivatives easier, and the study of such words has proven to be as useful as the study of the lexicon recorded in dictionaries (if not more) in defining and reorganizing theoretical paradigms (see Hathout, Montermini & Tanguy, 2008 for a discussion and an illustration of the consequences of the analysis of large corpora for

¹⁰ Of course, in covering all logical combinations we miss some obvious generalizations; for example it is well-known that Adj-N is a common merger, whilst Adj-V is not; moreover, some adjectival classes are more likely to merge with nouns; there are constraints on these mergers determined by the morphological type of the language, etc. In this case, our choice is a mere operational strategy. However, in encoding all possible combinations we produce a scheme which can in principle be applied to all languages.

theoretical morphology). In the discussion above, we outlined some of the main problems that a deep observation of real morphological data poses for purely “linear” models of morphology and we presented some alternatives to take the complexity of derivational processes into account. To simplify, current models of morphological analysis may be divided into linear (or additive) models and “relational” models. According to the first, the relation between a complex word and its base can be described as an oriented schema, in which some semantic instruction is added in concomitance with the adjunction of a phonological string. In the most extreme variants of such models (cf. Lieber, 1992), affixes and words are not different in nature: they are all lexical items, and the distinction between them lies in the selectional restriction attached to the elements of each category. Accordingly, the rules for the combination of morphemes into complex words are similar in nature to the rules for the combination of words into syntactic structures. Such models do not differ much from traditional structuralist models of morphemic combination. Their implementation into a method of automatic retrieval of affixed words in a database is, apparently, straightforward. However, these models present, as we have already seen, several serious drawbacks when they have to treat complex data, in particular data from fusional languages. These problems concern all the levels that are traditionally identified as intervening in a morphological relation, namely the formal (phonological), the categorial and the semantic level. From the point of view of phonology, for instance, a model that tries to draw a clear frontier between a base and an affix, viewed as separate but similar lexical objects, faces serious problems in dealing with allomorphy. Let us take, for instance, the deverbal action nouns of Italian in (12), for which we give the verbal base, at the infinitive and at the past participle, and the derived noun:

(12)	<i>educare</i>	‘educate’	<i>educato</i>	‘educated’	<i>educazione</i>	‘education’
	<i>esprimere</i>	‘express’	<i>espresso</i>	‘expressed’	<i>espressione</i>	‘expression’
	<i>esplosione</i>	‘explode’	<i>esplosione</i>	‘exploded’	<i>esplosione</i>	‘explosion’
	<i>immergere</i>	‘dip’	<i>immerso</i>	‘dipped’	<i>immersione</i>	‘immersion’

Defining what is the exact form of the suffix is not easy. The only common string for all the derivatives in question is *-ione*. However, this string can only be preceded by a reduced number of sequences, and in fact the suffix only appears, in complex words, under the forms *-zione* (phonologically [tsjone]) and *-(s)sione* (phonologically [sjone] or [zjone]). Moreover, the exact form that the affix has in the output depends, at least partially, on the form of the past participle. A purely additive model would be either obliged to miss this generalization by posing a maximally underspecified form for the affix (*-ione*), or to list all the possible

allomorphs and to describe, at the same time, the exact conditions for the choice of the appropriate allomorph.

On the other side, most of the models that we have informally labelled as “relational” assume that affixes, and the other exponent of morphological rules, should not necessarily be considered as lexical objects, or signs, but rather as the exponents of a morphological relationship. From this point of view, there is no difference between a segmental operation, like affixation, and a non-segmental operation, like, say, stress shift. Under this view, morphological rules, which emerge as patterns of correspondence between the words contained in the lexicon and are not reified into specific word-pieces (e.g. affixes). Thus, a rule is not a “recipe” for combining different lexical objects (e.g. a root and an affix), but an operation performed on a word in order to obtain another word. The affix is thus concomitant with the rule itself. Redundancy rules of this type were first proposed by Jackendoff (1975), and are, more or less explicitly, used in several current word-based models of morphology (cf. e.g. Bybee, 1985, 2007; Blevins, 2006; cf. also Booij’s *Construction Morphology*: Booij, 2010). A crucial question, in this respect, is the status of the lexicon. An (often implicit) assumption of additive approaches is that the best model minimizes stored information and gives priority to computation. Accordingly, the lexicon only contains idiosyncratic information and is an unstructured list of unanalysable items. However, there is a strong evidence, including psycholinguistic studies, that speakers do not necessarily store all and only irregular forms, and that regular forms, in particular the very frequent ones, may also be memorized (cf. Bybee, 2007). This is consistent with the fact that it is often impossible to establish in a deterministic way what is regular and what is idiosyncratic in the lexicon. This holds for phonology, as the examples in (12) show, but also for semantics (cf. Aronoff, 2007 for some observations on this matter).

Let us consider some other cases, and in particular such lexical gaps as the one exemplified in (13):

- (13) a. *estremo* *estremista* *estremismo* *estremizzare*
 ‘extreme’ ‘extremist’ ‘extremism’ ‘extremise’
 b. *catechista* *catechismo* *catechizzare*
 ‘catechist’ ‘catechism’ ‘catechise’

In a typical additive model the data in (13a) could be analysed by means of three separate WFRs, each one expressing a relation between the base (*estremo*) and a derivate. However, this sort of local relationship does not hold for other, similar, data, such as those in (13b), for which one would also like to establish a correspondence. Of course, an additive model could be modified in order to take such data into account, for instance by

identifying a form as the base (which could be a hypothetical bound form as **catec*, or one of the three derivatives listed), and by introducing rules of mutual deletion of the affixes involved. A relational model, on the other hand, simply identifies patterns of relations between the forms actually contained in the lexicon. Note that an advantage of this kind of model is that the rules they permit to identify can be viewed as non-oriented, which means that the very notion of “base” becomes a relative one. The fact that morphological rules are non-oriented is consistent, on the one side, with the idea that they are tools not only for the construction of new lexical items, but also for the analysis of the existing lexicon, and on the other side, with the empirical observation that, in real linguistic life, speakers are fortuitously exposed both to simplex and to complex words and that their morphological competence is precisely a means to manage the relation between the two. Of course, the connection between relational models and corpus searching techniques is less immediate than for additive models. The implementation of a model of this kind, in fact, cannot be realized by simply decomposing a string of characters and by giving a certain value to the substrings thus identified. The DG and the mechanism of encoding categorial relations we proposed in section 2, we argue, represent a first step towards the resolution of the problems connected with this issue.

4. CONCLUSIONS

In this paper we have presented and discussed some issues that emerge from a systematic interaction between morphology and computational linguistics. The way in which corpora are annotated and made searchable for complex morphological queries strongly depends on the theoretical framework assumed as reference point. In other words, any annotation of a corpus is the projection of a theory, and this is also true for morphology. Assuming – as we have done in this paper – that a single complex word can have more than one underlying structure or, in other words, that the order of application of a prefix and a suffix (if it is not supported by empirical evidence) can be a negligible criterion of analysis, is, by itself, a theoretical choice, and has consequences on the representation of affixes and word formation processes within the theoretical framework.

The problematic data illustrated in sections 1 and 3, which have often been considered as “marginal” or “odd”, seem in fact to be far more frequent and pervasive than one would think. Such cases and the practical problems they imply play against the “traditional” linear representation of word formation processes and additive models of morphology. The main challenge they pose concerns the elaboration of theoretical models of

word formation that put together the necessary internal consistency, a full accuracy in describing real and actual data, and a satisfactory operating capacity.

Our proposal of representing word formation processes through DGs is more than a mere “technical” solution for operating problems, since it has strong theoretical consequences. First of all, the DG assumes that it is not appropriate to represent affixes disregarding the whole morphological processes that govern their application. Consistently with the “relational” models mentioned above, we regard an affix as the exponent of a net of morphological relations or of grammatical frames; it specifies the information which is relevant for semantics and syntax. The initial steps in the derivation define just these frames; then bases (roots, stems, etc. depending on the morphological shape of each single language) are added to the representation. A graph, being just an abstract representation of a network of “objects”, where some pairs of objects are connected by directed edges, perfectly fits such situations. The degree of internal complexity of these nets can vary, according to different parameters. Among them, a crucial role is played by frequency and by the contexts of occurrence. The fact that in a DG the lexical categories (i.e. parts of speech) are indicated on the edge(s), and not explicitly on the affix(es) suggests that assigning a word class is a prerogative of the process as a whole, and not of the affix itself.¹¹ Moreover, it emphasizes the importance of context, since, as shown above, we have to admit that an important part of the information related to a complex word, including part of speech, is contextually determined. As it can vary according to usage, it cannot be pre-specified in the representation of an affix. Besides, the possibility of navigating the DG in more than one direction suggests that the link between complex morphological operations and a strict linear order of application of different affixes is less strong than one would think: morphological operations can create underlying structures that break the physical constraint on the linearity of linguistic signs;¹² this can make the issue of the relative order of application of affixes in complex words negligible.

Of course, our proposal should be tested on a wider amount of data;¹³ and its theoretical implications should be managed cautiously, but we are firmly convinced that the questions posed in this paper, concerning the relation between a theory of word formation and its operating consequences, represent a topic that no morphological theory can elude.

¹¹ This is true also for subcategorization frames, which we did not analyse in this paper.

¹² Cf. Moro (2006) for a similar situation in syntax.

¹³ It would be important to test this proposal on languages displaying non-concatenative morphology.

REFERENCES

- Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, MA: The MIT Press.
- Aronoff, M. (2007). In the beginning was the word. *Language* 83 (4), 803-830.
- Battista, M. & Pirrelli, V. (2000). *Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane*. Pisa: ILC-CNR.
- Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics* 42, 531-573.
- Booij, G. (2010). *Construction Morphology*. Oxford, Oxford University Press.
- Bybee, J. L. (1985). *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam/Philadelphia: John Benjamins.
- Bybee, J. L. (2007). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.
- Celata, C. & Bertinetto P. M. (2010). Per un'analisi morfologica del lessico italiano. *Quaderni del Laboratorio di Linguistica* 9 (1), 1-13.
- Delmonte, R. (2000). Parsing with GETARUN. *Proc.TALN2000, 7° Conférence annuel sur le TALN* (pp. 133-146). Lausanne.
- Hathout, N., Montermini, F. & Tanguy, L. (2008). Extensive data for morphology. using the World Wide Web. *Journal of French Language Studies* 18 (1), 67-85.
- Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon. *Language* 51 (3), 639-671.
- Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model for Word-form Recognition and Production*. Helsinki: University of Helsinki Department of General Linguistics.
- Lesmo, L. & Lombardo, V. (2002). Transformed subcategorization frames in chunk parsing. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)* (pp. 512-519). Las Palmas.
- Lieber, R. (1992). *Deconstructing Morphology. Word Formation in Syntactic Theory*. Chicago: The University of Chicago Press.
- Lindén, K., Silfverberg, M. & Pirinen, T. (2009). HFST tools for morphology. An efficient open-source package for construction of morphological analyzers. *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology 2009*. Zürich, Switzerland.
- Matthews, P. H. (1974). *Morphology*. Cambridge: Cambridge University Press.
- Moro, A. (2006). *I confini di Babele. Il cervello e il mistero delle lingue impossibili*. Milan: Bompiani.
- Pianta, E., Girardi, C. & Zanolì, R. (2008). The TextPro tool suite. *Proceedings of the 6th International Conference on Language Resources and Evaluation Conference (LREC2008)*. Marrakech, Morocco.
- Rossini Favretti, R., Tamburini, F. & De Santis, C. (2002). A corpus of written Italian: a defined and a dynamic model. In A. Wilson, P. Rayson & T. McEnery (Eds.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World* (pp. 27-38). Munich: Lincom-Europa.
- Scalise, S. (1994). *Morfologia*. Bologna: Il Mulino.
- Zanchetta, E. & Baroni, M. (2005). Morph-it! A free corpus-based morphological

resource for the Italian language. *Proceedings of Corpus Linguistics 2005*, Birmingham, UK: University of Birmingham.

Nicola Grandi

Dipartimento di Studi Linguistici e Orientali
Università di Bologna
Via Zamboni 33, I-40126 Bologna
Italy
e-mail: nicola.grandi@unibo.it

Fabio Montermini

CLLE-ERSS, CNRS & Université de Toulouse
Université de Toulouse le Mirail - Maison de la Recherche
5, allées Antonio Machado, F-31058 Toulouse
France
e-mail: fabio.montermini@univ-tlse2.fr

Fabio Tamburini

Dipartimento di Studi Linguistici e Orientali
Università di Bologna
Via Zamboni 33, I-40126 Bologna
Italy
e-mail: fabio.tamburini@unibo.it