Luigi Talamo, Chiara Celata

Toward a morphological analysis of the Italian lexicon: developing tools for a corpus-based approach

(versione provvisoria di un contributo sottoposto ad altra rivista)

1 Introduction

Few linguistic corpora provide morphological information: morphosyntactic features such as Part-of-Speech (POS) or lemma form are normally encoded in the corpus but information concerning the morphological structure, either inflectional or derivational, is hardly given.

Regarding Italian, one noticeable exception is *Morph-IT* (Zanchetta and Baroni 2005), a lexicon which contains the full paradigm of about 400,000 Italian verbs. Similar tools can conveniently be used for the automatic tagging of corpora.

Yet, for derivational morphology, no comparable instruments exist. The present project aims to fill this gap, with specific reference to derivation through affixation. The lexical database is CoLFIS (Bertinetto et al. 2005), a four million tokens corpus developed in the mid nineties with specific psycholinguistic purposes. For each complex form contained in CoLFIS, we will describe the affixes involved in the derivational processes as well as its formal and semantic features.[1]

The final aim is to realize a morphologically encoded corpus for Italian that allows for quantitative morphological studies on the language. Quantitative studies on Italian derivational morphology are at present almost impossible to realize because of the scarcity of reliable sources of distributional data. For instance, in their works on the productivity of Italian affixes, Gaeta and Ricca were not able to extract productivity values for the nominalizing suffix *-iere*, because the separation of the inflected form of *-iere* from those of *-iera* and *-iero* (which are two different nominalizing suffixes) "would have been far too much time-consuming" (Gaeta and Ricca 2003). Furthermore, because of the lack of morphologically annotated corpora, it is impossible to automatically distinguish a derivational affix from some other homographic string, as in *divertimento* 'entertaintment' ←*divertire* 'to entertain' + *-mento* '-ment' as opposed to *frumento* 'wheat'.

The structure of this paper is as follows. In §2 we provide a short description of CoLFIS. The criteria used to establish the data sample are presented in §3. The theoretical problems underlying morphological annotation will be discussed in §4. A detailed description of the annotation procedure is presented in §5, and §6 concludes.

[1]This project is part of a larger enterprise aimed at including information on both morphological and prosodic structure in CoLFIS. In a subsequent phase, we will undertake an analysis and annotation of the inflected forms in the corpus in addition to Part-Of-Speech tagging for the annotation of morphosyntatic categories. Lemmas are annotated prosodically with information about the position of lexical stress and syllabification. Support for this research comes from FIRB 2009-2012 "WIKIMEMO.IT: Il Portale della Lingua e della Cultura Italiana" (SNS research unit).

2 CoLFIS

The CoLFIS "Corpus e Lessico di Frequenza dell'Italiano Scritto" (Bertinetto et al. 2005) is composed of 3,798,275 lexical tokens sampled from a large corpus of Italian books, journals and newspapers. The corpus was originally selected on the basis of official statistical data on the reading preferences of Italian speakers provided by ISTAT (the national institute for demographic analyses) in 1993. CoLFIS was created with the specific purpose of representing the mental lexicon of the Italian speakers as reliably as possible (Laudanna et al. 1995: 104-106). As stated by the authors, CoLFIS was explicitly intended to be a tool for psycholinguistic research: the frequency of the occurrence of words is computed with respect to a carefully and purposely selected sample of texts that is expected to mirror the 'average' frequency of exposure of the individual lexemes by an 'average' Italian speaker. Thus, CoLFIS is representative of the lexicon that is concretely experienced by speakers in everyday reading (rather than actively produced in the spoken communication) of a large variety of commonly used written texts.

CoLFIS is structured with both a "Formario" (or list of lexical tokens) and a "Lemmario" (or list of lexical types), with the relevant frequency indices calculated with respect to the overall corpus and to its different subsections (according to the text typology). The "Lemmario" includes information on the grammatical category of the lemma.

It follows, then, that CoLFIS represents an attractive lexical database for corpus-based explorations of the properties of the Italian lexicon and its sub-structures (i.e., words' morphemic structure, derivational affixes). In particular, with respect to sub-lexical structures, quantitative studies on Italian are still at their very beginning (see §3 and Grossman and Rainer 2004 for an overview).

3 Annotating derivational morphology: the data sample

Derivational morphology plays a meaningful role in the definition of the Italian lexicon: in the *Grande Dizionario Italiano dell'Uso* (hencefort, GRADIT: GRADIT 2003), more than one third of the total lemmas (93,000 out of 250,000) is represented by derived words.

Concerning affixation, Italian has 91 prefixes (which forms roughly 17,000 lemmas) and 316 suffixes. These affixes can be classified according to the morphosyntactic category of their input i.e., verbalizing, adverbializing, nominalizing and adjectivalizing affixes, the last of which is the most numerous affixal class. Moreover, in showing a strong preference toward suffixation, Italian is consistent with a well-known typological claim (see, among others, Stump 2001:708-711).

Although the CoLFIS' lexicon is accessible in an alphabetical order, we started by analyzing and annotating complex forms whose affixes follow diversified criteria. We did not want to investigate high frequency affixes only: it is well-known that in the lexicon of a given language, a crucial role is played by certain mid-frequency affixes, which cover some specific lexical domains (Corbin's *rentabilité* - see also Gaeta and Ricca 2006:61). An example of this in Italian is the nominal suffix *-(t)ore*, e.g. *marcatore* 'soccer scorer'.

Therefore, on the basis of the available data on the frequency and productivity of Italian affixes, we set up a balanced affixes set of 40 affixes.

In Gaeta and Ricca 2006 the productivity values for a subset of derivational affixes were obtained using a modified version of Baayen's index of productivity P(N) (Baayen 2009, Baayen 1993), namely, the variable-corpus approach. This procedure involves the calculation of the ratio between hapax legomena and tokens for each affix but differs from the original index in that the ratio value is computed at equal token numbers for different affixes.

Some of the affixes selected for the purposes of this project were core derivational affixes, which are the affixes that are crucial for Italian word formation (Gaeta and Ricca 2003:89). Most of them are the following nominalizers:

- the suffix *-ezza*, which derives deadjectival quality nouns, e.g., *bellezza* ('beauty');
- the suffix *-ismo* which forms abstract nouns from nouns, verbs and adjective, e.g., *comunismo* ('communism'). This suffix is often found in paradigmatic alternation with suffix *-ista*, which is also a multi-inputs suffix yielding relational nouns and adjectives, e.g. *comunista* ('communist');
- the suffixes *-mento* and *-(z)ione* which form deverbal action nouns, e.g. *pagamento* ('payment'), *punizione* ('punishment');
- the suffix *-(t)ore*, which attaches to verbal bases, deriving agent and instrument nouns, e.g., *vaporizzatore* ('vaporizer').

Three suffixes form adjectives or attach to verbs:

- the adjectival suffix *-bile*, which form deverbal (and sometimes denominal) qualitative adjectives, e.g., *amabile* ('lovable');
- the negative prefix [1]*in-*, which attaches to adjectival bases, e.g., *instabile* ('unstable');
- the iterative prefix *re-*, which attaches to verbal bases, e.g., *rivendere* ('to resell').

Among the affixes listed by Gaeta and Ricca 2006 that are unproductive or that have limited productivity, we selected two nominal suffixes and one adjectival suffix:

- the suffix *-(z)a*, which forms deverbal/deadjectival action nouns, e.g., *decadenza* ('decay');
- the suffix *-izia*, which derives deadjectival quality nouns, e.g., *mestizia* ('sadness');
- the adjectival suffix *-evole*, which derives deverbal/denominal qualitative adjectives, e.g., *considerevole* ('valuable').

A different set of values is given by Gaeta and Ricca 2003:77 with logarithmic classes. Established by Thornton 1997, these classes order affixes by the natural logarithm of their absolute frequencies (i.e., the total number of tokens at the end of the corpus). Eight

classes are generally recognized, which range from class 2 (lowest frequency) to class 9 (highest frequency)[2].

From the highest classes (classes 6-9) we selected three suffixes:

- the nominal suffix *-iere* (seventh class), which forms denominal agentive nouns, e.g., *barbiere* ('barber');
- the adjectival (and sometimes denominal) suffix *-iero* (sixth class), which attaches to nominal bases and forms relational adjective (and sometimes agentive nouns), e.g., *alberghiero* ('daily');
- the adverbial suffix *-mente* (ninth class), which derives deadjectival adverbs, e.g., *lentamente* ('slowly').

From the lowest classes (classes 2-5) we selected seven affixes:

- the adjectival suffix *-aceo* (second class), which derives denominal qualitative adjectives, e.g., *conchigliaceo* ('shell-like');
- the pejorative suffix *-accio* (fourth class), which attaches to nominal bases, e.g., *poveraccio* ('poor man (pej.)');
- the collective denominal suffix *-aglia* (third class), e.g., *boscaglia* ('brush');
- the locative suffix [1]*-aio* (fourth class), whose inputs are nouns, e.g., *vespaio* ('wasps' nest');
- the denominal suffix [2]*-aio* (fifth class), which derives agentive nouns, e.g., *gelataio* ('ice-cream man');
- the nominal suffix *-igia* (second class), which forms quality nouns, e.g., *ingordigia* ('greed');
- the evaluative prefixes *mini-* and *micro-*, both denominal/deadjectival, e.g., *minigonna* ('miniskirt') and *micro-motore* ('micro-engine').

Lastly, we included a group of affixes whose productivity values are unknown, among which we had two denominal suffixes, two series of neoclassical suffixes, which are often found in paradigmatic alternation and attached to neoclassical bases, seven prefixes, mostly deverbal and/or deadjectival and two evaluative suffixes, which attach to nearly any bases:

- the agentive suffix *-ario* and its allomorph *-aro*: e.g., *bibliotecario* ('librarian');
- the ethnic suffix [2]*-ino*, e.g., *aretino* ('Aretinian');
- *-crazia*, *-crate* and *-cratico*, which form, respectively, abstract nouns, e.g., *burocrazia* ('bureaucracy'), relational nouns, e.g., *burocrate* ('bureaucrat') and relational adjectives, e.g., *burocratico* ('bureaucratic');

[2]Thornton does not explicity mention a first class: none of the affixes considered in her work display a lower frequency than class 2.

- *-logia* and *-logo*, forming abstract nouns, e.g., *biologia* ('biology') and relational nouns, e.g. *biologo* ('biologist'), respectively;
- the reflexive *auto-*, e.g., *auto-aiuto* ('self-help');
- the duplicative *bi-* and its allomorph *bis-*, e.g., *biunivoco* ('biunivocal');
- the directional [1]*in-*, e.g., *inaridire* ('to scorch');
- the negative *de-*, e.g., *de-industrializzazione* ('de-industrialization');
- the opposite [1]*s-*, e.g., *scoprire* ('to uncover') and *dis-*, e.g., *disobbedire* ('to disobey');
- intensive [2]*s-*, e.g., *scacciare* ('to chase away');
- the circumventive *trans-* and its allomorphs *tra-* and *tras-*, e.g., *transalpino* ('transalpine');
- the diminutive/ameliorative *-etto*, e.g., (*orsetto*) 'little bear' and [1]*-ino*, e.g., *trenino* ('little train').

4 Theoretical problems and criteria of analysis

A complex form is a word that has undergone some word formation processes. However, several linguistic facts, both diachronic and synchronic, may make this simple claim more complex. In our discussion we will mainly address derivation, though the theories discussed here aim to cover all kind of morphological processes.

As we were concerned with the task of identifying and annotating derived words in the Italian lexicon, the first problem that we encountered was the homography/homophony of some affixes with certain endings. Recall the example mentioned above of *diverti-mento* ('entertainment') vs. *frumento* ('wheat').

This is a clear-cut and well-defined case of distinction between a derived and a non-derived form, which can easily be accounted for by appealing to the diachronic origins of the relevant forms. A much more complex problem arises, though, when we have to evaluate in synchronic terms the morphological status of forms that have a diachronic account for their derivational nature, but that show different degrees of 'transparency' with respect to the original morphological process. For example, the derived form *osservazione* ('observation') is intuitively more transparent with respect to the word formation process that has attached *-(z)ione* to the verbal base *osserva-*, than *unzione* 'unction', whose base *unt-* (see below) is more difficult to recover in synchrony.

Therefore, these phenomena have to be treated in terms of graded notions of morphological transparency, which has to be further articulated in formal (i.e. morphotactic) and semantics component of the morphological process. A long-standing tradition of study has described morphological transparency with scales; this is the approach summarized in Dressler 2005, in which two functionalist scale, one for the morphotactic component and the other for the semantic transparency, are proposed.

In the following sections, we will review the fundamental tenets of Dressler and his colleagues' discussion on morphological transparency, and we will propose an adaption of the morphotactic and semantic scales for the purposes of the present study.

4.1 A scale of morphotactic transparency

Originally conceived in the framework of Natural Morphology, this scale classifies morphotactic transparency according to the degree of naturalness shown by phonological, morphological and morphonological phenomena involved in a morphological processes.

In his book on action nouns (Gaeta 2002), Gaeta adopts the Natural Morphology approach to describe four Italian derivational suffixes and shows that more transparent (i.e., more natural) suffixation processes also display higher values of productivity.

Following this procedure and according to Dressler's original scale (Dressler 1985:330-331), we propose the morphotactic scale represented in Table 1. The complex forms used in the examples are taken from CoLFIS.

DEGREE	NATURE OF PHENOMENON	EXAMPLE
I	none	*dichiara-* 'to declare (verbal theme)' + *-(z)ione* = *dichiarazione* 'declaration', *de-* + *tassare* = *detassare* 'to detax'
II	purely prosodic and phonological (e.g., resyllabification, assimilation)	sonorization: *[z]-debitare* 'to repay'
IV	morpho-phonological, without loss of morpho-phonological constituents (e.g., fusion, articulatory weakening)	affricativization: *unt-* 'to oil (irregular past participle)' → *un[ts]ione* 'unction'
V	morpho-phonological, with loss of morpho-phonological constituents (e.g., deletion)	*ipnotico* 'hypnotic' → *ipnot-izzare* 'to hypnotize'
VI	pure morphological (e.g., paradigmatic alternation of affixes)	*comunismo* 'communism' → *comunista* 'communist'
VII	lexical: weak suppletion	*pioggia* 'rain' → *pluv-iale* 'rain (adj.)'
VIII	lexical: strong suppletion	*guerra* 'war' → *bellico* 'war (adj.)'

Table 1: Scale of morphotactic transparency (after Dressler 1985).

Seven[3] degrees in Italian word formation processes are shown, and the level of morphotactic opacity increases as we move from degree I (where no intervening phonological processes obscure the relation between the base and the derived form) up to degree VIII.

The most transparent degree (I in the scale) is characterized by the absence of any intervening phenomena, resulting in the mere juxtaposition of the base form and the affix: e.g., *dichiara* ('to declare (verbal theme)') + *-(z)ione* → *dichiarazione* ('declaration'), *de-* + *tassare* 'to tax' → *detassare* 'to detax'. Pure phonological phenomena influence the morphotactics of complex forms classified under the II degree: for instance, the

[3] With respect to Dressler's 8-level original scale for English, one degree is missing in our proposal, namely, the degree that concerns the effects of neutralizing phonological rules, such as the flapping: *write* + *-er* → *wri*[ɾ] (this degree was originally conceived of as the third degree of morphotactic scale, occupying an intermediate position between the purely phonological and the morphological phenomena). This degree does not seem to pertain to Italian morphotactics. Moreover, since several Italian suffixes begin with a vowel, most suffixation processes entail resyllabification, as in *ri.ci.cla.re* ('to recycle') → *ri.ci.clag.gio* ('recycle (noun)'): thus, at least in principle, we have to reclassify these morphological processes under degree II because of their prosodic nature.

sonorization of the prefix *s-* caused by the assimilation with a following voiced consonant, *s-* + *debitare* → *[z]debitare* 'to repay'.

Degrees IV and V are both of a morpho-phonological nature, featuring phonological phenomena that affect morphological constituents. For degree IV, we included complex forms whose morphotactics does not cause the loss of morphemes; degree V concerns more opaque morphotactic processes involving the loss of morphemes. For example, under degree IV, we classified the derivational process in *unt-* 'to oil (irregular past participle)' + *-ione* (allomorph of *-(z)ione*), which involves the fusion of the voiceless dental stop /t/ with the palatal glide /j/: **un[tj]one* → *un[ts]ione* ('unction'). An example for degree V is *ipnotic-* 'hypnotic' + *-izzare*, which involves the deletion of the segment /ic/ and results in *ipnotizzare* ('to hypnotize') (**ipnot[ic]izzare*).

The morphotactics of degree VI forms is the result of purely morphological phenomena, such as the paradigmatic alternation of suffixes in *comun-ismo* ('communism') / *comun-ista* ('communist'). Finally, lexical phenomena are grouped under the most opaque degrees: weak suppletion processes (VII degree), e.g. *pioggia* ('rain') / *pluv-iale* ('rain (adj)') and strong suppletion processes (VIII degree), e.g. *guerra* ('war') / *bellico* ('war (adj)').

Indeed, (native) speakers are not generally aware of word formation processes that are influenced by prosodic and phonological phenomena (degree II). Their 'morphotactic awareness' starts to appear when derivation is influenced by morphophonological or morphological phenomena (degress IV to VI), eventually becoming a 'lexical necessity' with degress VII and VIII. It is therefore possible to propose three macro-classes of morphotactic transparency: a first macro-class including the degrees I and II of the scale, which contain phonological phenomena; a second macro-class, encompassing degrees IV to VI, which contain morphological phenomena; finally, a third macro-class, of the degrees VII and VIII, which contain the lexical phenomenas.

The role of allomorphy

One point that we want to stress here is that the morphotactic transparency scale does not account for allomorphy, neither of the base nor of the affix. In principle, we have to specify in the annotation which allomorph of the base and/or the affix is involved in the formation of a given complex form, and then describe the morphotactics accordingly. For instance, the complex form *unzione* not only shows a certain degree of morphotactic opacity (degree IV) but is also based on an irregular past participle, *unto*.

In this respect, deadjectival and denominal derivation does not display particular problems: on the contrary, the forms of some deverbal complex words, such as those with *-(z)ione*, *-(t)ore*, *-(t)orio* and *-(t)ura*, do show allomorphy. According to Gaeta and Ricca 2006:75-78, the base form used for these suffixation processes may be represented by three distinct verbal forms: the verbal theme (i.e., the verbal root plus the thematic vowel), the irregular past participle and the Latinate past participle. Furthermore, the suffixes themselves display a certain degree of allomorphy in that the verbal theme requires the suffix forms *-zione*, *-tivo*, *-tore*, *-torio* and *-tura*, whereas the past participle requires *-ione*, *-ivo*, *-ore*, *-orio* and *-ura*[4].

[4]According to Rainer 2001:388 and fn.7-8, Latinate past participle is also the base for some suffixations

We exemplify this issue with *-(z)ione*'s case, which is previously discussed in Thornton 1990-1991 and Scalise 1984. For a similar discussion on the other suffixes and their relationship, which is partly paradigmatical as Rainer 2001 suggests, to suffix *-(z)ione*, see Rainer 2001:386-389 and fn.8.

The verbal theme, as proposed by Thornton 1990-1991, accounts only for complex forms such as *contraddizione* ('contradiction'), and fails to capture derived words such as *discussione* ('discussion') (cf. **discutezione*).

In contrast, the irregular past participle, as suggested by Scalise 1984:67, correctly accounts for *discussione* but fails to predict *contraddizione* (as it generates **contraddettione*). Finally, as shown by Gaeta and Ricca 2006:76, both approaches cannot explain some forms based on the Latinate perfect participle (Lat.p.p.), such as *adesione* ('adhesion') (Lat.p.p. *adesus*): Thornton's approach gives **aderizione*, and Scalise's proposal **aderitione*.

Once more, such a wide array of allomorphy can easily be captured with the parameter of Natural Morphology's naturalness. Indeed, the most natural base is one that obeys certain criteria of morphotactic and morphosemantic transparency. This kind of base is represented in Italian derivation by the bare lexical root, which is used in denominal affixation (e.g., *muscol-* 'muscle' in *muscoloso* 'muscular') because it is similar to the autonomous word in its uninflected form. Italian composition actually makes use of this form, e.g., *capo-branco* ('leader of the pack'). (Dressler 2005:273)

From this perspective, the Italian deverbal word formation process through affixation appears less natural because the most common base that it uses is the verbal theme, as in *ama-* ('to love (v.t.)') →*amabile* ('lovable'). Other base forms are less natural and are restricted to certain suffixes (see above): the irregular Italian past participle, e.g., *unzione* ('unction') and the Latinate past participle, e.g., *sensibile* ('sensitive'). See Gaeta and Ricca 2006:79 for data supporting these claims.

Returning to the Italian denominal derivation, unnatural bases are those produced by suppletion, as in *addomin-* for *addome* ('abdomen') in *addominale* 'abdominal'. Finally, non-autonomous bases used in base-less derivation, e.g. *vulnera-* in *vulnerabile* 'vulnerable' are probably the least natural bases.

A hierarchy of naturalness can be established for allomorphs as well, with the most common allomorph being the most natural one. It is often the case that the most natural base combines with the most natural allomorph of the affix. Compare, for example, *ama-* ('to love (v.t.))' →*ama-bile* ('lovable') with *vis-* ('to view (Lat.p.p.))' →*vis-ibile* 'viewable'. The allomorph *-ibile* is less productive and therefore less natural than *-bile* in derivations.

4.2 A scale of morphosemantic transparency

Turning now to the semantic aspect of transparency, we again adhere to the parameter of naturalness, as proposed by Dressler 2005. Regarding the semantics of word formation processes, the most appropriate term to be used, as Dressler himself suggests, is 'morphosemantic transparency'.

---

in *-bile*: *condivisibile* ←Lat.p.p. *condivisus*.

A morphosemantic scale is described by Dressler 2005:275 for English compounding and it is based on the psycholinguistic model by Libben 1998. On the basis of solid psycholinguistic evidences, Libben argued for a three-level model of representing and processing compounds. At the first level (the 'stimulus' level), a morphological parser processes compounds to feed the lexicon with morphemes. On the second level (the 'lexicon' level), compounds get their structure: transparent and semi-opaque compounds, such as *door-bell* and *strawberry*, respectively, display (lexical) connections among their constituents whereas pseudocompounds, such as *hum-bug*, do not. To account for the differences between transparent and semi-opaque form, which are represented in the same way at the lexicon level, one has to postulate a third level (the 'conceptual' level) which covers the semantics of constituents. At this level, the constituents of *door-bell* are both analyzed as transparent, whereas *straw* in *strawberry* is analyzed as opaque. (Libben 1998:35-39)

Libben's model is transposed in Dressler's (2005) morphosemantic scale in four levels, organized according to the semantic transparency of the constituents. The highest (most transparent) level is represented by compounds that show transparency in both head and dependent. A lower (i.e., more opaque) level includes compounds such as English *strawberry*, which only shows transparency in the head. The opposite case, (i.e., those compounds with opaque head and transparent dependent, as in English *jail-bird*), belongs to the third level. Finally, compounds such as English *hum-bug* are opaque in both constituents and are therefore classified under the fourth, most opaque level.

As Dressler suggests, this morphosemantic scale can be extended to other mechanisms of word formation. According to this view, we propose an adaption of the scale for the purposes of derivational morphology through affixation, as shown in Table 2.

The most transparent morphosemantic degree in composition has, as Dressler 2005:272 points out, a clear counterpart in affixation: by showing transparency in both the base and the affix, derived forms such as *divertimento* ('entertainment') or *stappare* ('uncork') display the same morphosemantic structure as a compound such as *door-bell*.

Following the assumption that affixes are always the heads in derived words, semi-opaque compounds with transparent head such as *strawberry* may be seen as similar to derived words such as *potabile* ('drinkable') or *restaurare* ('restore') (level 3 in Table 2). Previously labelled by Corbin 1987:187 as '*mot complexes non construits*' and discussed for Italian by Gaeta and Ricca 2003:71 who call them 'base-less forms', these derived forms show transparency in their word formation meaning (*Wortbildungsbedeutung*) but opacity in the meaning of their base because **pota-* and **staurare* are not lexical morphemes in Italian.

In fact, if a 'strawberry' is 'a sort of berry' for English speakers, *potabile* is perceived by Italian speakers as 'something which can be X-ed', according to the meaning of the suffix (and in spite of the unanalyzable base). In addition, the suffix clearly signals the morphosyntactic category of the derived word (e.g., *potabile* is an adjective). Finally, as claimed by Gaeta and Ricca 2003:71, these base-less formations seem to have, at least to a certain extent, a psycholinguistic weight, which parallels the behaviour of English opaque compounds as demonstrated by Libben 1998:32-35.

The *strawberry* type shows similarities with respect to another group of morphosemantically opaque derived words, namely, the words pertaining to level 2 of Table 2. These

forms have undergone some lexicalization processes: according to Dressler 2005:271, these processes have their endpoint in fossilization. We have tried to capture this continuum by splitting this level into two sub-levels: sub-level 2b contains quasi-fossilized forms (i.e., forms that have almost lost their word formation meaning) and sub-level 2a illustrates derived forms whose bases have almost lost their transparency, while at the same time retaining their *Wortbildundgsbedeutung*. For instance, *costituzione* ('constitution (law)') is no more a *nomina actionis* as it was at the time of its formation and *disinvolto* ('self-assured') does not mean 'unconstrained' whereas *aquilone* ('kite') mantains the property of referring to 'something big', although not exactly to 'a big eagle' (Italian *aquila* 'eagle'), and *disintegrare* maintains the negative meaning of the affix, though is no more longer related to the meaning of the verb *integrare* ('integrate').

LEVEL	EXAMPLE	TRANSPARENCY	
		BASE	AFFIX
1	*divertimento* 'entertainment', *stappare* 'uncork'	+	+
2a	*aquilone* 'kite', *disintegrare* 'disintegrate'	±	+
2b	*costituzione* 'constitution (law)', *disinvolto* 'self-assured'	±	±
3	*potabile* 'drinkable', *restaurare* 'restore'	-	+

Table 2: Three levels of morphosemantic transparency for Italian affixed forms: a '+' stands for 'full transparency', '-' stands for full opacity, and '±' stands for partial transparency as discussed in the text.

Even if it is not adopted in Dressler's original scale, the semantic feature of compositionality, present in Libben's model, allows us to account for differences between the sub-levels 2a and 2b. According to Libben 1998, the absence of compositionality characterizes exocentric compounds (i.e., those compounds whose head, although transparent, is not an hyperonym of the resulting compound, as in English *yellow-belly*, which is not a 'belly' but a 'coward'). Thus, derived forms pertaining to sub-level 2b, such as *costituzione*, are similar to non-compositional compounds whereas derived forms relating to sub-level 2a, such as *aquilone*, show a compositional meaning similar to *strawberry*-like compounds.

There is not any parallel degree in affixation for the *jail-bird* type in compounding. It is possible that other morphological processes of word formation, such as zero-affixation or conversion, (e.g., *bacio* 'kiss' $\rightarrow$ *baciare* 'to kiss'), could be interpreted as cases of derivation with 'non-transparent head'.

4.3 Process ordering in word formation

Italian derivational morphology can be described as 'layered' (see, among others, Manova and Aronoff 2010:113-114), in that process ordering in word formation is governed by semantics. Thus, describing the order in which several derivational processes apply on a given base is a crucial component of the analysis. As an intuitive example, there is a clear distinction between forms derived with *-izzare* + *-bile*, such as *privatizzabile* ('something

that can be privatized') and forms derived in *-bile* + *-izzare*, such as *sensibilizzare* ('to sensitize').

Affixation may occur in conjunction with conversion and (to a certain extent) compounding, as shown, for instance, by the action noun *facilitazione* ('easing'), which is derived through three different word formation processes: [[[[facil]ità](c)]zione]. The base *facile* ('easy') is first suffixed by *-ità*, then converted (as signalled by (c)) into the verb *facilitare* ('to ease') and finally subject to a second suffixation with *-zione*. Finally, there are complex forms that are derived through simultaneous processes; this case is generally defined as parasynthetic or multiple-affixation. The latter term is probably inaccurate for Italian word formation because most parasynthetic processes actually consist of affixation in addition to conversion, as in *s-* + *cardine* 'hinge' → *scardinare* 'unhinge' (but **cardinare* and **scardine*).

Similarly to what Bybee 1985 has shown for inflection (i.e., morphemes that are closer to the base tend to encode inherent morphosyntatic properties), we can be predict that inner word formation processes are morphotactically and morphosemantically opaque than outer word formation processes.

Take for instance the two double-suffixed forms *responsabilizzare* ('to make someone aware') and *sensibilizzare* ('to make someone sensitive'). Both display an inner suffixation process that is more opaque than the outer one. Indeed, the derived form *responsabile* ('aware') has to be classified under the third morphosemantic level (the base form **responsa* is not a lexical morpheme in Italian), whereas *sensibile* ('sensitive') pertains to level 2a (because of having undergone a semantic drift, its base form *sentire* means 'to feel').

5 Analysis and annotation of some derived forms

5.1 Analysis

The analysis of some complex forms according to the criteria above is presented in Table 3, in which both scales of morphotactic and morphosemantic transparency are shown. Since Natural Morphology claims there is a "tendency towards iconicity between morphosemantic and morphotactic transparency/opacity" (Dressler 2005:273), we expect that our analysis will provide interesting data on this assumed correlation between the two scales.

For example, some derivational processes are characterized by an iconic relationship between the semantics of the derivation and the formal means employed by the language to achieve it. Italian diminutives are a well-known case of this relationship because they often involve a process of consonant palatalization or vowel raising which can be seen as a process of sound iconism related to the 'diminutive' or 'emotive' sense of /i/ vowel and /i/-like sounds (Merlini Barbaresi and Dressler 1994).

As discussed in subsection a., allomorphy has also to be included in the annotation. The allomorphic alternations of the bases have been indicated in Table 3:

- 'root': the base form is the lexical bare root;
- 'v.t.': the base form is the verbal theme;

Example	Morphotactic transparency	Morphosemantic transparency			Allomorphy	
	DEGREE	LEVEL	BASE	AFFIX	BASE	AFFIX
muscoloso 'muscular'	I	1	+	+	root	-oso
amabile 'lovable'	I	1	+	+	v.t.	-bile
visibile 'viewable'	I	1	+	+	Lat.p.p.	-ibile
sdebitare 'to repay'	II	1	+	+	v.t.	s-
unzione 'unction'	IV	1	+	+	irr.It.p.p.	-ione
sensibile 'sensitive'	I	2a	±	+	Lat.p.p.	-ibile
ipnotizzare 'to hypnotize'	V	1	+	+	root	-izzare
fascismo 'fascism'	VI	1	+	+	root	-ismo
addominale 'abdominal'	VII	2a	±	+	suppl.	-ale
bellico 'war (adj)'	VIII	2a	±	+	suppl.	-ico
svaligiare 'to ransack'	II	2a	±	+	root	s-
costituzione 'constitution (law)'	I	2b	±	±	Lat.p.p.	-zione
vulnerabile 'vulnerable'	VIII	3	-	+	baseless	-bile

Table 3: Analysis of morphotactic and morphosemantic transparency and processes of allomorphy for some Italian complex words.

- 'Lat.p.p.': the base form is the Latinate perfect participle;
- 'It.irr.p.p.': the base form is the Italian irregular past participle;
- 'suppl.': the base form is a suppletive form;
- 'baseless': the base form is a non-autonomous lexical morpheme (base-less derived forms).

It seems that even allomorphy is iconically reflected on the morphotactic and the morphosemantic scales. For instance, *unzione*, which is based on the Italian irregular past participle *unt-* with suffix *-ione*, is morphotactically opaque; similarly, *possibile*, which is morphosemantically opaque, is formed with the Latinate perfect participle *poss-* and the suffix *-ibile*.

It is likely that a thorough investigation of the semantics of derivation in Italian will uncover many other aspects of this iconic relationship, thereby allowing a test of the naturalness prediction in a statistically reliable way.

5.2 Annotation

The structure of the CoLFIS "Lemmario" (see §2) currently available consists of a simple database with sixteen fields for each lexical entry, which primarily contain quantitative data such as the lemma's absolute and relative frequencies as distinguished by source (books, newspapers, journals), its orthographic length and grammatical category.

The morphological annotation here maintains this basic configuration, organizing the new informations in a database whose structure closely resembles the internal structure of a complex word. For each complex lemma, the base and the (either unique or multiple) word formation process(es) (wfp) involved are specified in the relevant fields (Figure 1):

lemma	base	wfp1	wfp2	wfp3	wfp4	wfp5	wfp6

Figure 1

For each derived lemma, six slots are provided and are filled according to the order of occurrence of the relevant processes:

lemma	base	wfp1	wfp2	wfp3	wfp4	wfp5	wfp6
stappare	tappo	Cnv	1S				
impermeabilizzare	permeare	IN-P	BILE-P	IZZARE			

Figure 2

In the example of Figure 2, the label `Cnv` indicates a conversion noun-verb (*tappare* 'to cork') and `1s` signals the affixation with [1]*s-*, a prefix with a negative meaning (similar to English 'un-'). Parasynthetic process are marked by the trailing character -P on each wfp involved.

The word-formation process fields contain the morphological annotation in the following form:

```
AFFIX:allomorphy:morphotactic transparency+morphosemantic transparency
```

This is similar to the annotation proposed by Zanchetta and Baroni 2005 for the project Morph-IT!, which aims to describe the Italian inflectional morphology. For instance, the Italian inflected verb *canteremo* 'we will sing' is annotated as follows:

```
canteremo cantare VER:ind+fut+1+p
```

We substitute the label `VER`, indicating the Part-of-Speech tagging, with the label `AFFIX`, which corresponds to the affix. As a general criterion, we decided to use the commonest/most natural affix as label (see above). For instance the suffix *-bile* in *visibile* is:

```
BILE:ibile
```

As for the base, we note lexical allomorphy, using the lemma form as a label; for example, the base in *visibile* is:

```
VEDERE:latpp
```

with `latpp` belonging to the features discussed above and means Latinate Perfect Participle.

As for transparency, the values used in the two scales are labelled as follows:

Scale of morphotactic transparency

- first degree: `mt1`;
- second degree: `mt2`;

- fourth degree: `mt4`;
- fifth degree: `mt5`;
- sixth degree `mt6`;
- seventh degree `mt7`;
- eighth degree `mt8`.

Scale of morphosemantic transparency

- transparent level: level 1 (label `ms1`);
- lexicalization level: level 2a and 2b (label `ms2a` and `ms2b`, respectively);
- base-less level: level 3 (label `ms3`).

We can now translate some complex forms into the metalanguage in Figure 3.

lemma	base	wfp1	wfp2
muscoloso	MUSCOLO:root	OSO:oso:mt1+ms1	
amabile	AMARE:vt	BILE:bile:mt1+ms1	
visibile	VEDERE:latpp	BILE:ibile:mt1+ms1	
stappare	TAPPO:root	Cnv	1S:s:mt1+ms1
unzione	UNGERE:irrpp	ZIONE:ione:mt4+ms1	
scardinare	CARDINE:root	Cnv-P	1S:s:mt1+ms1-P
sensibile	SENTIRE:latpp	BILE:ibile:mt1+ms2	
vulnerabile	VULNERA:baseless	BILE:bile:mt8+ms3	
facilitare	FACILE:root	ITÀ:ità:mt1+ms1	Cnv
responsabilizzare	RESPONSA:baseless	BILE:bile:mt8+ms3	IZZARE:izzare:mt1+ms1
addominale	ADDOME:suppl	ALE:ale:mt7+ms2a	
bellicoso	GUERRA:suppl	ICO:ico:mt8+ms2a	OSO:oso:mt1+ms1

Figure 3: Annotation of some complex forms.

This annotation is developed to achieve the maximum compatibility with other computational tools, and as such it will be possible to automatically tag any corpus by means of the morphological information provided by the annotation of CoLFIS.

6 Conclusion

We have illustrated that the annotation procedures for Italian affixation processes have clear theoretical implications and that they require thorough investigation into the fine-grained aspects of derivational phenomena, which have to be classified with a limited set of clear-cut categories.

We have noted the two criteria of morphotactic and morphosemantic transparency to define a grid of interpretable features for the variety of attested forms in derivation. We believe that the final product will allow for testing some fundamental predictions of Natural Morphology on affix-base combination constraints through the quantitative assessment of different patterns of morphotactic-morphosemantic iconicity.

Furthermore, our CoLFIS annotation will promote experimental investigations (both corpus-based and psycholinguistic) on several aspects of the morphology of Italian.

Bibliographical References

Baayen, H. (1993). "On frequency, Transparency and Productivity." In: *Yearbook of Morphology 1992*. Ed. by G. Booij and J. Van Marle. Dordrecht: Kluwer, pp. 181–208.

Baayen, H. (2009). "Corpus linguistics in morphology: morphological productivity". In: *Corpus Linguistics. An international handbook.* Ed. by A. Luedeling and M. Kyto. Mouton De Gruyter: Berlin, pp. 900–919.

Bertinetto, Pier Marco et al. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. URL: `http://linguistica.sns.it/CoLFIS/Home.htm`.

Bybee, Joan (1985). *Morphology: a study of the relation between meaning and form.* Amsterdam: John Benjamins.

Corbin, Danielle (1987). *Morphologie dérivationelle et structuration du lexique*. Tübingen, Niemeyer.

Dressler, W. U. (1985). "On the Predictiveness of Natural Morphology". In: *Journal of Linguistics* 21.2, pp. 321–337.

Dressler, W. U. (2005). "Word-formation in Natural Morphology." In: *Handbook of Word-Formation (Studies in Natural Language and Linguistics Theory, volume 64)*. Ed. by P. Stekauer and R. Lieber. Springer, pp. 335–352.

Gaeta, L. and D. Ricca (2003). "Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data". In: *Italian Journal of Linguistics/Rivista di linguistica* 15.1, pp. 63–98.

Gaeta, L. and D. Ricca (2006). "Productivity in Italian word formation: a variable-corpus approach". In: *Linguistics* 44.1, pp. 57–91.

Gaeta, Livio (2002). *Quando i verbi compaiono come nomi: un saggio di Morfologia Naturale*. Milano: Franco Angeli.

GRADIT (2003). *Grande dizionario italiano dell'uso, edizione su CD-ROM*. Ideato e diretto da Tullio De Mauro, Torino: UTET.

Grossman, Maria and Franz Rainer, eds. (2004). *La formazione delle parole in italiano*. Tübingen: Niemeyer.

Laudanna, A. et al. (1995). "Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente." In: *II Giornate internazionali di Analisi Statistica dei Dati Testuali.* Ed. by S. Bolasco, L. Lebart, and A. Salem. Roma: Cisu, pp. 103–109.

Libben, G. (1998). "Semantic Transparency in the Processing of Compounds: Consequences for Representation, Processing, and Impairment". In: *Brain and Language* 61, pp. 30–44.

Manova, S. and M. Aronoff (2010). "Modeling affix order." In: *Morphology* 20.1, pp. 109–131.

Merlini Barbaresi, Lavinia and Wolfgang Ulrich Dressler (1994). *Morphopragmatics: diminutives and intensifiers in Italian, German, and other languages*. Berlin/New York: Mouton de Gruyter.

Rainer, F. (2001). "Compositionality and paradigmatically determined allomorphy in Italian wordformation." In: *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday.* Ed. by Ch. Schaner-Wolles, J. H. Rennison, and F. Neubarth. Rosenberg & Sellier: Torino., pp. 900–919.

Scalise, Sergio (1984). *Generative Morphology*. Dordrecht: Foris.

Stump, G. (2001). "Affix position." In: *Language Typology and Language Universals: An International Handbook*. Ed. by M. Haspelmath et al. De Gruyter: Berlin-New York., pp. 708–714.

Thornton, A. M. (1990-1991). "Sui deverbali italiani in -mento e -zione (I-II)". In: *Archivio glottologico italiano* 75 and 76, 169–207 and 79–102.

Thornton, A. M. (1997). "Quali suffissi nel "vocabolario di base"?" In: *A limiti del linguaggio*. Ed. by F. Albano Leoni et al. Bari: Laterza, pp. 385–396.

Zanchetta, E. and M. Baroni (2005). "Morph-it! A free corpus-based morphological resource for the Italian language". In: *Corpus Linguistics 2005* 1.1. ISSN: 1747-9398.