# How to deal with morphotactic and morphosemantic transparency in the morphological annotation of an Italian corpus

Luigi Talamo & Chiara Celata (luigi.talamo|chiara.celata@sns.it)
Laboratorio di Linguistica - Scuola Normale Superiore, Pisa

## 1 - Overview

**Background**: Few linguistic corpora provide morphological information, either inflectional or derivational. Regarding Italian, we may rely on Morph-IT (Zanchetta and Baroni 2005), containing the full paradigm of about 400,000 Italian verbs, and the ongoing project AnIta, a morphological analyser based on 120,000 lemma (Grandi, Montermini, and Tamburini 2011).

**Aim of the project**: To realize a database of morphologically annotated Italian derived forms that allows for quantitative studies of the dynamics of derivation.

**Corpus**: 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 describe the formal and semantic aspects of the derivational process.

**Dataset**: 40 Italian affixes (11 prefixes, 29 suffixes), covering approximately 10,000 types in COLFIS (details in Talamo and Celata 2011).

**Annotation features**: (i) base/derivational morpheme; (ii) type of base/affix allomorph, (iii) morphotactic transparency; (iv) morphosemantic transparency.

**Theoretical and empirical issues**: Derivational processes have to be treated in terms of a graded notion of morphological transparency, which has to be further analyzed in its formal (i.e. morphotactic) and semantic aspects. In this project, we follow a long-standing tradition of study that describes morphological transparency with scales (Dressler 2005), and propose an adaptation for Italian.

## 2 - Morphotactic transparency

One of the key assumptions of Natural Morphology is that formal complexity of word forms has consequences on the cognitive level of linguistic processing. The table below is an adaptation of the morphotactic scale discussed in Dressler 2005 to Italian derivational processes. The fourth column shows the annotation labels used in this project.

| DEGREE | NATURE OF PHENOMENON | EXAMPLE | LABEL |
|---|---|---|---|
| I | none | *de-* + *tassare* = *detassare* 'to detax' | mt1 |
| II | purely prosodic and phonological (e.g., resyllabification, assimilation) | sonorization: *[z]-debitare* 'to repay' | mt2 |
| 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' | mt4 |
| V | morpho-phonological, with loss of morpho-phonological constituents (e.g., deletion) | *polemico* 'polemical' → *polem-izzare* 'to polemize' | mt5 |
| VI | pure morphological (e.g., paradigmatic alternation of affixes) | *comunismo* 'communism' → *comunista* 'communist' | mt6 |
| VII | lexical: weak suppletion | *pioggia* 'rain' → *pluv-iale* 'rain (adj.)' | mt7 |
| VIII | lexical: strong suppletion | *guerra* 'war' → *bellico* 'war (adj.)' | mt8 |

Seven 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.

With respect to Dressler's 8-level original scale for English, one degree (III) is missing in our proposal, namely, the degree that concerns the effects of neutralizing phonological rules (e.g., flapping: write + -er >wri[ɾ]er). 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.

According to Natural Morphology, forms belonging to the most natural degrees of the scale are expected to show higher values of productivity. At the end of the project, we will be able to verify this hypothesis over a large amount of data.

## 3 - Morphosemantic transparency

The following scale is inspired by Libben's (1998) four-degree scale of morphosemantic transparency in English compounding. The fourth column shows the annotation labels used in this project.

| LEVEL | EXAMPLE | TRANSPARENCY | | LABEL |
|---|---|---|---|---|
| | | BASE | AFFIX | |
| 1 | *stappare* 'uncork', *allenamento* 'training' | + | + | ms1 |
| 2a | *aquilone* 'kite', *disintegrare* 'disintegrate' | ± | + | ms2a |
| 2b | *costituzione* 'constitutional law', *intrattenere* 'to entertain' | ± | ± | ms2b |
| 3 | *sbiadire* 'unfade', *potabile* 'drinkable' | - | + | ms3 |

At Level 1, both the base and the affix show full morphosemantic transparency. Derived forms that are opacified due to a process of lexicalization (half-transparency: ±) but still retain their affix meaning are assigned to Level 2a. Level 2b shows half-transparency of the affix instead. Finally, 'base-less' forms (Gaeta and Ricca 2003:71) that show transparency in their word formation meaning but opacity in the meaning of their base because the base is not a lexical morpheme in Italian, are assigned to Level 3.

## 4 - Annotation

Derived forms are described according to their base and the word formation processes (**wfps**) synchronically recoverable. Apart from few exceptions, we assume that wpfs are linear and the structure of the derived word is layered (Manova and Aronoff 2010:113-114). Seven different slots are provided, one for the annotation of the base and up to six for the wfps. Parasynthetic process are marked by the trailing character -P on each wfp involved.

| Derived | Base | Wfp1 | Wfp2 | Wfp3 | Wfp4 | Wfp5 | Wfp6 |
|---|---|---|---|---|---|---|---|
| AFFABULAZIONE | FAVOLA:suppl | AD:ad:mt7:ms2a-P | C:N_V-P | ZIONE:zione:mt1:ms1 | | | |
| ANTIPROIBIZIONISTA | PROIBIRE:vt | ZIONE:zione:mt1:ms1 | ISMO:ismo:mt1:ms1 | ANTI:anti:mt1:ms1 | ISTA:ista:mt6:ms1 | | |
| ASSENTEISMO | BASELESS:unrecoverable | NZA:nza:mt8:ms3 | NTE:nte:mt6:ms1 | ISMO:ismo:mt1:ms1 | | | |
| COSTITUZIONALISTA | COSTITUIRE:latpp | ZIONE:ione:mt4:ms2b | ALE:ale:mt1:ms1 | ISTA:ista:mt1:ms1 | | | |

## 5 - Further development

- ✓ Evaluation: Inter-annotation agreement (in progress)
- ✓ Psycholinguistic tests to assess the validity of morphosemantic scale
- ✓ Quantitative assessment of frequency and productivity for Italian derivational patterns
- ✓ Quantitative investigation of aspects of morphotactic-morphosemantic iconicity
- ✓ Automatically tagging of Italian corpora with morphological information

## References - 1


- Pier Marco Bertinetto et al. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. URL: http://linguistica.sns.it/CoLFIS/Home.htm.
- W. U. Dressler (2005). "Word-formation in Natural Morphology." In: *Handbook of Word-Formation (Studies in Natural Language and Linguistics Theory, volume 64)*. Ed. by P. Štekauer and R. Lieber. Springer, pp. 335–352.
- L. Gaeta 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.
- N. Grandi, F. Montermini, and F. Tamburini (2011). "Annotating large corpora for studying Italian derivational morphology". In: *Lingue e Linguaggio* 2.10, pp. 227–244.
- G. Libben (1998). "Semantic Transparency in the Processing of Compounds: Consequences for Representation, Processing, and Impairment". In: *Brain and Language* 61, pp. 30–44.


## References - 2


- S. Manova and M. Aronoff (2010). "Modeling affix order." In: *Morphology* 20.1, pp. 109–131.
- L. Talamo and C. Celata (2011). "Toward a morphological analysis of the Italian lexicon: developing tools for a corpus-based approach". In: *Quaderni del Laboratorio di Linguistica* 1.10. URL: http://linguistica.sns.it/QLL/QLL11.htm.
- E. Zanchetta and M. Baroni (2005). "Morph-it! A free corpus-based morphological resource for the Italian language". In: *Corpus Linguistics 2005* 1.1.