On the significance of the corpus size in affix-order research[i]

Stela Manova & Luigi Talamo

Abstract

This article discusses suffix ordering in derivation in a Slavic language (Bulgarian) and a Romance language (Italian) and examines the reliability of different sources of data. The theoretical part is couched in a cognitive approach to affix order (Manova 2011b) which sees derivational suffix combinations as binary structures of the type SUFF1-SUFF2[ii] where SUFF1 has three valency positions for further suffixation: $SUFF2_N$, $SUFF2_A$ and $SUFF2_V$. There is either a single SUFF2 of each lexical category or if more than one SUFF2 of the same lexical category is available, there is one SUFF2 that attaches by default, that is, the majority of the types are derived by a single SUFF2, or the available SUFF2 suffixes express completely different semantics (e.g., an abstract noun and an object). The data come from various sources, including specialized electronic resources and corpora. A specialized resource (one annotated for research on derivational morphology) based on a well-balanced relatively small corpus appears as reliable as a one-hundred-times-larger electronic corpus.

Keywords: *affix ordering, derivation, corpus size, Bulgarian, Italian, Cognitive approach, lexical category, fixed and predictable suffix combinations*

1. Introduction

This article tackles affix ordering in derivation with a focus on suffixation and analyzes data from a Slavic language (Bulgarian) and a Romance language (Italian). Our goal is to answer the question of what sources of data can be used for investigation of suffix ordering in derivation in an inflecting language, that is, in a language in which derivational suffixes are not always word-final but followed by inflection. Nevertheless, our research results are not language-specific but comparable with research on affix ordering in other languages, especially with research on affixation in English, the most studied language with respect to affix ordering. For English, many resources for research on suffixation are available but in this language derivational suffixes are always word-final. Actually, most search tools in dictionaries and corpora allow search for only word-initial and word-final segments. As in order to be comparable cross-linguistically a study should be situated in a theoretical framework, we follow an approach that has been tested successfully against data from Bulgarian, English and Russian, the so-called Cognitive approach to affix order (Manova 2011b, 2015a).

The two languages under scrutiny in this paper, Bulgarian and Italian, exhibit a very similar morphological organization: they have a relatively simple noun inflection, but the verb inflection is complex; the majority of their derivational suffixes are nominalizing and there are only a few verbalizing suffixes. Put differently, both languages' morphology relies on lexical categories such as nouns, adjectives, and verbs. Additionally, in both languages there is a clear distinction between derivational and inflectional suffix slots

(Skalička 1979, Manova 2011a), on the one hand, and between purely derivational, i.e., non-evaluative, and evaluative suffix slots, on the other hand (Manova 2010, 2011b). This is illustrated in (1) with an example from Bulgarian and in (2) with an example from Italian:

(1) *ogništencata* 'the small fireplaces' (Bulgarian)
ogn -išt(e) -enc(e) -a -ta
fire derivational suff diminutive suff plural infl definite article

(2) *cinturini* 'little belts, wristbands' (Italian)
cint -ur(a) -in(o) -i
(to) hold derivational suff diminutive suff plural infl

As shown with the examples (1) and (2), the purely derivational suffixes are the closest to the root, they may be followed by evaluative suffixes such as diminutive and augmentative ones after which comes the inflection. We call suffixes such as *-isht(e)* and *-ur(a)* purely derivational and set them apart from the evaluative suffixes such as *-enc(e)* and *-in(o)* and the inflectional suffixes (the plural marker *-a* and the definite article *-ta* in Bulgarian and the plural inlfection *-i* in Italian) not only because of their position in the word form but because the three types of affixes also exhibit different affix-ordering peculiarities (Manova 2010, 2015a): purely derivational suffixes can form mirror image combinations of the type AB–BA, see (4a) for Bulgarian and (4b) for Italian. Additional examples of mirror-image combinations from Bulgarian can be found in Manova (2010, 2015a), and from Italian in Talamo (2015).

(4) Mirror image combinations: AB–BA
(4a) *-(l)iv*+*-ost* versus *-ost*+*-(l)iv* (Bulgarian)
săn-liv-ost 'sleepiness' but *mil-ost-iv* 'merciful'
(4b) *-egg(iare)* + *-evol(e)* versus *-evol(e)* + *-egg(iare)* (Italian)
man-egg-evole 'handy' but *piac-evol-eggiare* 'to behave in a pleasing manner'

In contrast to the purely derivational suffixes, evaluative suffixes can be repeated, that is, they may form combinations of the type AA where two or more diminutive suffixes follow each other, examples in (5). On the patterns of the Bulgarian evaluative suffixes, see Manova & Winternitz (2011), and on Italian evaluative suffixes, Merlini Barbaresi (2012).

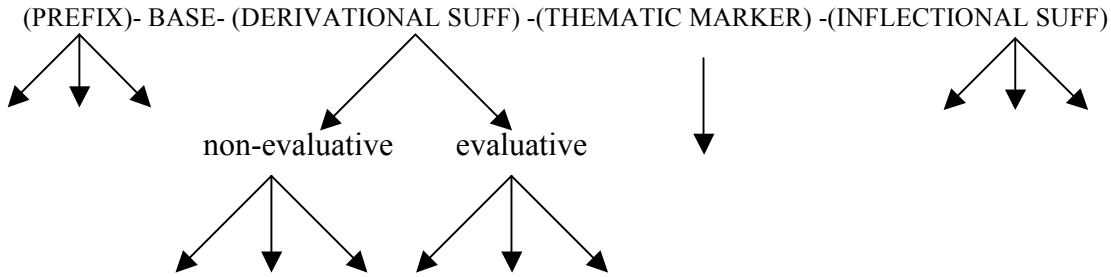
(5) Repetition of diminutive suffixes: AA
(5a) *det-enc-ence* 'child-DIM-DIM' (Bulgarian)
(5b) *fett-in-ina* 'slice-DIM-DIM' (Italian)

The inflectional suffixes differ from both the purely derivational and the evaluative suffixes in the sense that they do not exhibit any of the above-illustrated peculiarities, i.e., inflectional suffixes neither form AB–BA permutations (recall the examples in (4)) nor can be repeated (recall (5)). Actually, the inflectional suffixes always follow a fixed templatic order (Manova 2010 for examples from Bulgarian). As the purely derivational suffixes are the greatest number of the three types of suffixes and their combinability is least restricted, their behavior is most difficult to explain. Therefore in the present article, we focus on the combinability of

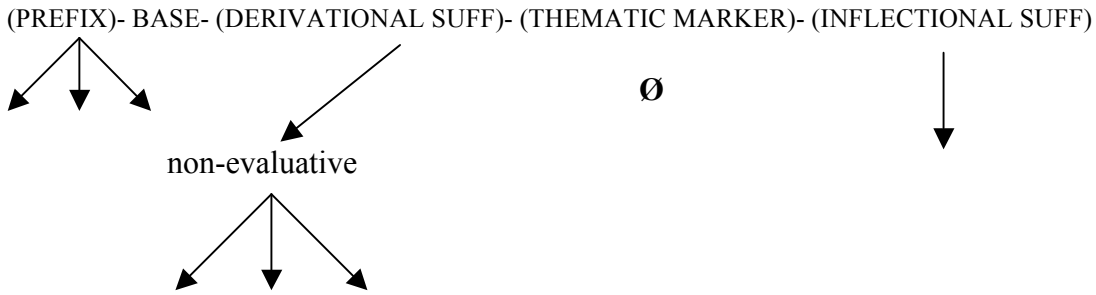
exactly this type of suffixes.

Based on the above observations, in this paper we follow a domain-specific approach (similar to that in Manova 2010), i.e. we assume that different types of rules are responsible for the ordering of the suffixes in the three domains, the purely derivational one, the evaluative one and the inflectional one, that is, the different types of suffixes should be analyzed differently with respect to affix order. As already mentioned, we tackle only the behavior of the purely derivational suffixes and set the latter apart from the evaluative and the inflectional suffixes. Such an approach is also in line with research on the ordering of the English derivational suffixes. The English evaluative suffixes cannot be repeated and they are, as a rule, treated together with the non-evaluative suffixes. As for the English inflectional suffixes, it has been assumed in the literature on affix order that they are not relevant to the ordering of the derivational suffixes, see, e.g., the most recent approach to affix ordering in English, the so-called Complexity-Based Ordering, Hay & Plag 2004, Plag & Baayen 2009, among others). The two schemas of the structures of the Bulgarian and English words from Manova (2011b) illustrate these observations:

(6) Affix order domains in the structure of the Bulgarian word



(7) Affix order domains in the structure of the English word



In (6) and (7), every slot and sublsot that can host more than one affix is associated with more than one arrow, that is: a single arrow means that within a word, only one single affix can occur in that slot; two arrows stand for two (types of) affixes; and three arrows mean that more than two affixes can co-occur in a particular slot. Ø in (7) indicates that there are no thematic markers in English, at least English does not possess affixes that could be seen as parallel to the thematic markers in other languages. The evaluative suffixes in English behave like the purely derivational suffixes and are therefore placed in the slot of the latter. For the

ordering of the prefixes which is outide the sope of this paper, see the discussion in Manova (2015a).

If the evaluative and the inflectional suffixes in a language are a fairly limited number and it is possible to list all combinations of the evaluative suffixes with one another and all combinations of the inflectional suffixes in a language (Manova 2015a), the purely derivational suffixes are a larger number and it is hard, if not impossible, to investigate all their combinations in a language. Therefore, a study on affix ordering in derivation usually discusses a set of suffixes and either combines those suffixes with one another (see, e.g., analyses that follow the Parsability Hypothesis or Complexity-Based Ordering, Hay 2003, Hay & Plag 2004, Plag & Baayen 2009, Talamo 2015, among others) or tries to list all the suffixes in a language that can follow the suffixes from a set under investigation (Aronoff & Fuhrhop 2001, Manova 2011, 2015a). The latter strategy is also used in stratal approaches (Siegel 1974; Allen 1978; Selkirk 1982; Kiparsky 1982, Mohanan 1986; Giegerich 1999) and in approaches that rely on selectional restrictions such as Fabb (1988) and Plag (1996, 1999). On this issue, see also the explanations of the approaches to affix order in Manova & Aronoff (2010) and Manova (2014).

In the present article, we examine two sets of derivational suffixes, one from Bulgarian and one from Italian, and consider all combinations of those suffixes with all other suffixes in the respective language. As we try to consider all combinations of a given suffix with all other suffixes, a question about the source(s) from which those combinations should be extracted arises. In the discussion we pay special attention to the answer of this question. As for the available sources of data for the languages under scrutiny, Bulgarian and Italian, there is no electronic corpus annotated for research on derivational morphology in Bulgarian, while there are a few electronic resources that can be used for that purpose in Italian (see the discussion in Talamo and Celata 2011 and Talamo et al., in press). We claim that a relatively small reverse dictionary is a good starting point for research on suffix order in derivation in an inflecting language and that a specialized electronic resource based on a well-balanced small corpus makes the same predictions as a very large corpus.

Finally, the research reported herein can be seen as being generally in line with Štekauer (1998 and later work) where it is claimed that word-formation is an independent component, interrelated with the lexical component, though we do not differentiate a conceptual level, a semantic level and an onomasiological level but assume following Cognitive semantics (e.g. Fillmore's 1982 Frame semantics) and Conceptual semantics (Jackendoff 1990) that there is no principle difference between meaning and conceptualization. Ourr approach is also a strictly synchronic one, thus we, in contrast to Štekaer's analysis, do not consider any diachronic observations.

The article has the following structure. The next section 2 sets up the theoretical scene. In section 3 the sources of data and the sets of suffixes are introduced. Section 4 accommodates the discussion and in section 5 conclusions are drawn.

2. Theoretical framework

Our research is couched within a recently suggested Cognitive approach to affix ordering (Manova 2011b). The approach assumes that general cognitive principles are also operative in grammar (Langacker 1987, 1991; Taylor 2002; and Geeraerts 2006) and that lexical categories such as noun, adjective and verb and semantic categories such as person, object,

etc. are cognitive in nature. So far, the Cognitive approach has been tested against data from Bulgarian and English (Manova 2011b) and Russian (Manova 2015a). In contrast to other approaches that treat all suffixes that can follow a particular SUFF1 together (Table 1 with data from Bulgarian and Table 2 with examples from Italian), the Cognitive approach assumes that a derivational SUFF1 has three valency positions for further derivation, i.e., the derivational SUFF2 suffixes that can immediately follow SUFF1 in a word form are distributed into three groups according to their lexical-category specifications: $SUFF2_N$, $SUFF2_A$ and $SUFF2_V$ (Table 3 and Table 4).

Table 1. Combinability of the Bulgarian suffix *-ar*$_N$ (based on Manova 2011b)

SUFF1	**Lexical category of SUFF1**	**SUFF2**	**Examples**	**Translation**
-ar	N person	*-stvo*, *-ski*	*aptek-ar-stvo* *aptek-ar-ski*	all pharmacists / being a pharmacist pharmacist's

For readers unfamiliar with Bulgarian, *aptek-ar* 'pharmacist', the base of the two examples in Table 1 (*aptek-ar-stvo* and *aptek-ar-ski*) is derived from *aptek-a* 'pharmacy' where *-a* is an inflectional suffix.

Table 2. Combinability of the Italian suffix *-izzare*$_V$ (based on *la Repubblica* corpus)

SUFF1	**Lexical category of SUFF1**	**SUFF2**	**Examples**	**Translations**
-izzare	V caus	*-mento, -zione, -tore, -bile, -(t)orio*	*volgarizzamento* *americanizzazione* *potabilizzatore* *utilizzabile* *privatizzatorio*	popularization americanization water purifier usable privatizatory

Table 3 below is a version of Table 1 (data from Bulgarian) and Table 4 is based on Table 2 (data from Italian). In tables 3 and 4, in contrast to tables 1 and 2, SUFF2 suffixes are classified according to their lexical-category specifications.

Table 3. Combinability of the Bulgarian suffix -*ar*$_N$ (SUFF2 classified for lexical category)

SUFF1	Lexical category of SUFF1	SUFF2	Examples	Translations	SUFF2 suffixes of the same lexical category in numbers
-ar	N person	N: *-stvo*	*aptek-ar-stvo*	all pharmacists / being a pharmacist	**N: 1**
		A: *-ski*	*aptek-ar-ski*	pharmacist's[iii]	**A: 1**

Table 4. Combinability of the Italian suffix -*izzare*$_V$ (SUFF2 classified for lexical category)

SUFF1	Lexical category of SUFF1	SUFF2	Examples	Translations	SUFF2 suffixes of the same lexical category in numbers
-izzare	V caus	N: *-mento* (4), *-zione* (>1000), *-tore* (>150)	*volgarizzamento americanizzazione potabilizzatore*	popularization americanization water purifier	**N: *3***
		A: *-bile* (>100), *-(t)orio* (10)	*utilizzabile privatizzatorio*	usable privatizatory	**A: *2***

As can be seen in Table 3, the Bulgarian SUFF1 -*ar*$_N$ combines with only one SUFF2$_N$, *-stvo*, and with only one SUFF2$_A$, *-ski*. We call combinations such as *-ar-stvo* and *-ar-ski* ***fixed***. In the tables below, fixed combinations are marked by a bold unit, i.e., by **1**, as it is done in the last column in table 3. The Italian suffix -*izzare*$_V$ (Table 4) combines with more SUFF2 suffixes than the Bulgarian *-ar*, namely with three nominalizing suffixes, -*mento*$_N$, -*zione*$_N$, and -*tore*$_N$, and with two adjectivizing suffixes, -*bile*$_A$ and -*(t)orio*$_A$. The numbers in brackets after the SUFF2 suffixes in Table 4 indicate the numbers of types derived with the respective SUFF2, i.e., -*bile*$_A$ derives more than 100 types while -*(t)orio*$_A$ derives only 10 types. In such cases, we speak of suffixation by default and the default suffix is the one that derives the majority of the types, in our case this is the suffix -*bile*$_A$. As the types derived by -*(t)orio*$_A$ are a very limited number, we assume that they should be rote-learned. Based on the data analyzed (35 suffixes from each of the two languages under investigation, Bulgarian and Italian) and the fact that in all cases of suffixation by default the suffixes that compete with the default suffix derive up to 10 types each, we postulate that a default suffix derives more than 10 types, while a combination that derives up to 10 types is rote-learned. The same observation has been made for Russian (Manova 2015a). We think that even if there is a rule that derives the up-to-ten types with SUFF2 suffixes that compete with the default SUFF2 for the SUFF1, the mere fact that those types are less than 10 requires the speaker to know the respective words by heart. Of the nominalizing combinations, -*mento*$_N$ and -*zione*$_N$ derive abstract nouns, -*zione*$_N$ being the default suffix (derives over 1,000 types), and -*tore*$_N$ derives objects. Thus, -*tore*$_N$ does not compete with -*mento*$_N$ and -*zione*$_N$ for the base suffix -*izzare*$_V$ but is assigned based on intentional semantics, that is, on what the speaker wishes to say. We

classify combinations such as *-izzare+-zione* (default combination) and *-izzare+-tore* (semantically determined combination) as predictable combinations. As already mentioned, we understand combinations such as *-izzare+-mento*, which derive less than 10 types, as rote-learned. Therefore, all three nominalizing suffix combinations in Table 4 are classified as predictable. ***Predictable combinations*** are in bold italic in Table 4 and the other tables below.

The theoretical framework of this study mixes principles and assumptions from Cognitive grammar (Langacker 1987, 1991; Taylor 2002; and Geeraerts 2006), including recent research in Cognitive neuroscience (Mestres-Missé et al. 2010, and the references therein), and Natural morphology (Dressler et al. 1987; Dressler 2005). The approach is defined as cognitive because with Cognitive linguistics (Langacker 1987, 1991, Taylor 2002 & Geeraerts 2006) it is assumed that grammar is an inventory of units (phonological, semantic, or symbolic structure) that have been established, or entrenched, in the speaker's mind through (frequency of) previous use. As typical of a cognitive account, the approach is usage-based and the following types of relations are of particular importance for the analysis: the whole-part relation; the schema-instance relation and the similarity-identity relation. With cognitive linguistics, it is also assumed that all aspects of cognition are shaped by aspects of the body. Roughly, we experience the world through our senses and general cognitive principles are also operative in linguistics.

With the traditional approaches to WF, we presume that a suffix tends to combine with suffixes of lexical categories different from its own, that is, that WF is prototypically word-class changing (Dressler 1989). However, as already mentioned, this research goes further in arguing that there usually exists only one combination with a suffix of a particular lexical category, Manova (2011b). Manova (2011b) sees the lexical-category specification of a suffix as definable on the basis of cognitive knowledge, which is similar to how cognitive linguists such as Langacker (1987) and Croft (2001) define N, A and V. Langacker (1987), based on *relationality* (i.e., +/- relational) and *way of scanning* (whether summarily scanned, i.e., conceived statistically and holistically, or sequentially scanned, i.e., mentally scanned through time), recognizes *things* (N), *processes* (V), and *modifiers* (ADJ). Croft (2001) defines objects, properties, and actions in terms of four semantic properties: *relationality, stativity, transitoriness*, and *gradability*. Thus prototypically, nouns name things or objects, verbs denote processes or actions, and adjectives are modifiers and express properties.

Additionally, we understand research in Cognitive neuroscience showing that nouns and verbs have different representations in the brain (Mestres-Missé et al. 2010, among many others) as supportive for the correctness of an approach that pays attention to the lexical category specification of an affix.

With respect to the cognitive nature of the semantic categories used in the analysis, following Cognitive semantics (e.g., Fillmore's 1982 frame semantics) and Conceptual semantics (Jackendoff 1990), we assume that there is no principle difference between meaning and conceptualization.

We also refer to Natural morphology, a semiotic and cognitively oriented theory of morphology compatible with cognitive grammar (Dressler 1990). According to the naturalness parameter of iconicity (constructional diagrammaticity), there are different types of affixation (see also Manova 2011a): affixation by addition in which addition of meaning is reflected by addition of form; and the less iconic affixation by substitution (truncation in Aronoff 1976). The English derivation *play-ful* → *play-ful-ness* and its Bulgarian equivalent *igr-iv* → *igr-iv-ost* are examples of affixation by addition, while the derivation *Marx-ism* → *Marx-ist* is an instance of affixation by substitution. Although *playfulness* / *igrivost* and

Marxist are analyzable as compositional units, only affixation by addition involves affix ordering. Thus in this study we will always control how two suffixes interact, and make a clear distinction between affixation by addition and by substitution.

In sum, the proposed research is based on the following assumptions: a) suffix combinations are pieces of structure entrenched in the speakers mind; b) they are best analysed in terms of binary combinations (SUFF1-SUFF2), the direction of derivation being from SUFF1 to SUFF2; c) suffixes are lexical entries (i.e. word-like) and the lexical category (N, V, A) and the semantics (e.g., persons, objects, etc.) of a suffix govern that suffix's combinability; d) SUFF1 is usually followed by a single suffix of a particular lexical category (we call such combinations *fixed*); e) if a particular SUFF1 is followed by more than one SUFF2 of the same lexical category: there is either SUFF2 that applies by default (i.e. the majority of the derivatives exhibit that suffix); or the SUFF2 semantics helps speakers differentiate among the different options for SUFF2 (we refer to combinations such as those in (e) as *predictable*.)

As the number of the existing suffix combinations and the number of the types derived by a particular SUFF2 are important parameters in our approach, we, in order to reduce omissions and challenge our theoretical assumptions as much as possible, looked for data in two languages, Bulgarian and Italian. The electronic recourses available for research on affix ordering (word-formation) in Italian do not exist for Bulgarian and the data from the latter language come primarily from a reverse-dictionary of a small size. Thus, it is no surprise that all combinations of the Bulgarian derivational suffixes reported in the literature, see Manova (2010, 2011b), are fixed and predictable. Therefore, it was important to test our approach against data from a language such as Italian for which also electronic resources for research on derivational morphology are available.

3. Data

There are two electronic corpora of Bulgarian, the *Bulgarian National Corpus* and the *Bulgarian Reference Corpus*. However, neither corpus is annotated for research on derivational morphology. Additionally, as derivational suffixes in an inflecting language such as Bulgarian are not word-final but followed by inflection, i.e., they are word-internal, the use of electronic corpora for research on affix order is very problematic, since the corpora search tools are, as a rule, not designed for search of word internal segments. The situation for Italian is different and, as already mentioned, there are a few specialized resources, databases and corpora, annotated with information on derivational affixation (for a discussion on this issue in relation to Italian, see Talamo and Celata 2011, Talamo et al., in press).

3.1. Bulgarian

The Bulgarian data discussed in this paper come from the 1975 *Reverse dictionary of Bulgarian,* containing a bit over 70,000 words. We preferred the older edition of the reverse dictionary as the more recent one (Murdarov et al. 2011) lists a lesser number of lexemes, about 65.000 words. Table 5 contains a sample of the Bulgarian data analyzed in Manova (2011b) and illustrates the theoretical assumptions explained in the previous section. As can be seen, all combinations in Bulgarian are either fixed or predictable. Actually, in Table 5 there is only one instance when a suffix combines with two suffixes of the same lexical

category, see suffix *-ina* (number 6 in table 5). In this case, the SUFF2 *-en*$_A$ derives 36 types while its competitor *-ski*$_A$ forms only 9 types. Thus, *-in-en*$_A$ is the default combination.

Table 5. Combinability of Bulgarian suffixes (sample, based on Manova 2011b)

No	SUFF$_1$	Lexical & semantic category of SUFF$_1$	SUFF$_2$ according to its lexical category	Examples of SUFF$_1$-SUFF$_2$ combinations	Translations	SUFF2 suffixes of the same lexical category in numbers
1.	***-(it)ba***	N verbal	N: *-ar* A: *-en*	*svat-b-ar* *svat-b-en*	wedding-guest wedding-	**N: 1** **A: 1**
2.	***-(iz)acija***	N abstract	A: *-onen*	*privat-izaci-onen,*	privatization-	**A: 1**
3.	***-(n)ica***	N location	N: *-ar* A: *-en*	*voden-ič-ar* *mel-nič-en*	watermiller mill-	**N: 1** **A: 1**
4.	***-ar***	N person	N: *-stvo* A: *-ski*	*aptek-ar-stvo* *aptek-ar-ski*	all pharmacists pharmacist's	**N: 1** **A: 1**
5.	***-ec***	N person	N: *-estvo* A: *-ki/-eski*	*tvor-č-estvo* *bor-č-eski*	artwork, creativity fighting	**N: 1** **A: 1**
6.	***-ina***	N location	A: *-en* (36), *-ski* (9)	*ravn-in-en* *plan-in-ski*	plain- mountain-	**A: *2***
7.	***-at***	A qualit	N: -ost	*ust-at-ost*	talkativeness	**N: 1**
8.	***-est***	A qualit	N: *-ost*	*por-est-ost*	being porous	**N: 1**
9.	***-iča***	V IMPFV inchoat	N: *-(V)ne*	*vazhn-ič-ene*	airs and graces	**N: 1**
10.	***-ira***	V IMPFV durat	N: *-(V)ne*	*pilot-ira-ne*	piloting	**N: 1**

3.2. Italian

The Italian data come from an annotated lexicon specialized for research on derivational morphology, the so-called *derIvaTario* (Talamo et al., in press) and from a large corpus of Italian, *la Repubblica* corpus (Baroni et al. 2004).[iv]

DerIvaTario is an annotated lexicon of the Italian derivates and contains over 11,000 entries. It was developed at Scuola Normale Superiore (SNS) in Pisa and is based on another resource, the *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*, Bertinetto et al. (2005). CoLFIS is a carefully balanced corpus of written Italian that contains over 3 M tokens and is meant to represent the mental lexicon of the ideal Italian speaker, more precisely reader (Laudanna et al. 1995), as the corpus is sampled from a variety of Italian books, journals and newspapers. It is designed on the basis of the official statistical data for the reading preferences of the Italians as provided by ISTAT (the national institute for demographic analysis) in 1993. CoLFIS has a number of special features. However, we will not pay special attention to those features as they are not directly relevant to our research. For information on CoLFIS, we refer the curious reader to Laudanna et al. (1995), and to the CoLFIS website: http://linguistica.sns.it/CoLFIS/Home.htm. The *derIvaTario* lexicon features morphological segmentation of derivatives, information on stem and affix allomorphy, as well as morphotactic and morphosemantic analyses of each word-formation step. Based on *derIvaTario*, we analyzed the combinations of 35 derivational suffixes in

Italian and they all confirm the cognitive approach (Manova 2011b) followed in this paper, i.e., all Italian suffix combinations are, like the Bulgarian ones in table 5, fixed and predictable. Table 6 contains a sample of our Italian data. More specific are only the combinations of the suffix *-izzare*$_V$ (number 8 in table 6) but we already explained them in section 2.

Table 6. Combinability of Italian suffixes (sample, based on *derIvaTario*)

No	SUFF1	Lexical& semantic category of SUFF1	SUFF2	Examples	Translations	SUFF2 suffixes of the same lexical category in numbers
1.	***-ese***	A rel	N: -ità (2) -ismo (2) V: -izzare (1)	*torinesità* *francesismo* *francesizzare*	the essence of being Turinese gallicism frenchify	**N: *2*** **V: 1**
2.	***-evole***	A qualit	N: -ezza (8), -ismo (1) V: -izzare (1)	*confortevolezza* *colpevolismo* *colpevolizzare*	comfortableness assumption of guiltiness to make sb feel guilty	**N: *2*** **V: 1**
3.	***-ico***	A rel	N: -ità (>10), -ismo (3) A: -oso (1) V: -izzare (3)	*classicità* *romanticismo* *bellicoso* *pubblicizzare*	classical antiquity romanticism warmongering advertise	**N: *2*** **A:1** **V:3**
4.	***-ile***	A rel	N: -ità (2), -ismo (3)	*signorilità,* *maschilismo*	class, elegance sexism	**N: *2***
5.	***-ino***	A rel	N -ismo (2), -ità (1)	*alpinismo* *latinità*	mountaineering classical antiquity	**N: *2***
6.	***-ismo***	N abstr	Ø	Ø	Ø	Ø
7.	***-ista***	N pers	Ø	Ø	Ø	Ø
8.	***-izzare***	V caus	N: -mento (1), -zione(>100) -tore (21) A: -bile (8) -(t)orio (1)	*volgarizzamento* *americanizzazione* *potabilizzatore* *utilizzabile* *privatizzatorio*	popularization americanization water purifier usable privatizatory	**N: *3*** **A: *2***
9.	***-oso***	A qualit	N -ità (>10), -ismo (3), -ario (locat) (1)	*faticosità* *virtuosismo* *lebbrosario*	laboriousness virtuosity leper colony	**N: *3***
10.	***-(t)orio***	A rel	Ø	Ø	Ø	Ø

la Repubblica corpus is a very large corpus (approximately 330 M tokens) and contains texts from Italian newspapers. It is tokenized, pos-tagged, lemmatized and categorized in terms of

genre and topic, but there is no annotation for derivational morphology. Nevertheless, since, in comparison to *la Repubblica*, CoLFIS is a very small corpus, we manually checked the combinations of the 35 Italian suffixes we investigated with the help of *derIvaTario* in *la Repubblica*. Table 7 contains the suffixes from Table 6 (those from *derIvaTario*) but now described according to their occurrences in *la Repubblica*. In the next section we compare Table 6 and Table 7 and discuss the differences between the pieces of information on affix combinability provided by *derIvaTario* and *la Repubblica*. Suffixes that are closing in *derIvaTario* and *la Repubblica* (see numbers 6, 7, and 10 in table 6 and table 7) were additionally checked for possible combinations on the Internet. In the discussion in the next section, we pay attention to this issue, too.

Table 7: Combinability of Italian suffixes (sample, based on *la Repubblica*)

	SUFF1	**SUFF1 lexical category & semantics**	**SUFF2**	**Examples**	**Translations**	**SUFF2 suffixes of the same lexical category in numbers**
1.	***-ese***	A rel	N: -ità (24) -ismo (8) -eria (6) V: -izzare	*torinesità* *francesismo* *giapponeseria* *francesizzare*	the essence of being Turinese gallicism collection of Japanese objects frenchify	**N: *3*** **V: 1**
2.	***-evole***	A qualit	N -ezza (8), -ismo (1) V -izzare (1)	*confortevolezza,* *colpevolismo* *colpevolizzare*	comfortableness assumption of guiltiness to make sb feel guilty	**N: *2*** **V: 1**
3.	***-ico***	A rel	N -ità (>10), -ismo (>10**)** A -oso (1) V -izzare (3)	*classicità* *romanticismo* *bellicoso* *pubblicizzare*	classical antiquity romanticism warmonger advertise	**N: *2*** **A: 1** **V: 1**
4.	***-ile***	A rel	N -ità (5), -ismo (6)	*signorilità* *maschilismo*	class, elegance sexism	**N: *2***
5.	***-ino***	A rel	N -ismo (>10), -ità (2)	*alpinismo* *latinità*	mountaineering classical antiquity	**N: *2***
6.	***-ismo***	N abstr	Ø	Ø	Ø	Ø
7.	***-ista***	N pers	Ø	Ø	Ø	Ø
8.	***-izzare***	V caus	N: -mento (4), -zione (>1000) -tore (>150) A: -bile (>100) (default) -(t)orio (10)	*volgarizzamento* *americanizzazione* *potabilizzatore* *utilizzabile* *privatizzatorio*	popularization Americanization water purifier usable privatizatory	**N: *3*** **A: *2***
9.	***-oso***	A qualit	N -ità (>10), -ismo (4), -ario (locat) (1)	*faticosità, virtuosismo* *lebbrosario*	laboriousness virtuosity leper colony	**N: *3***
10.	***-(t)orio***	A rel	Ø	Ø		Ø

4. Discussion

Table 8 below contains the data from both Table 6 (based on *derIvaTario*) and Table 7 (based on *la Repubblica*). As can be seen from Table 8, *derIvaTario* and *la Repubblica* differ in the number of types of some of the combinations, see especially the suffixes 1, 3 and 8 in Table 8. As could be expected in most cases, *la Repubblica* has more types than the *derIvaTario* which is no surprise given the almost one-hundred-times-smaller size of CoLFIS on which the *derIvaTario* is based but *la Repubblica* and *derIvaTario* coincide with respect to fixed and predictable combinations, see the last two columns in Table 8.

Table 8. *DerIvaTario* versus *La Repubblica*

No	SUFF1	SUFF1 lexical category & semantics	SUFF2	derIvaTario	La Repubblica
1.	***-ese***	A rel	N: -ità (2)**(24:rep)** (default for nouns, derives quality nouns) -ismo (2) **(8:rep)** (closing, derives abstract nouns) **-eria (6:rep)** (derives abstract nouns and objects) V: -izzare (1) **(>10:rep)**	**N: *2*** **V: 1**	**N: *3*** **V: 1**
2.	***-evole***	A qualit	N -ezza (8), -ismo (1) V -izzare (1)	**N: *2*** **V: 1**	**N: *2*** **V: 1**
3.	***-ico***	A rel	N -ità (>10), -ismo (3)**(>10:rep)** (see the explanations in 1. -ese) A -oso (1) V -izzare (3)	**N: *2*** **A: 1** **V: 1**	**N: *2*** **A: 1** **V: 1**
4.	***-ile***	A rel	N -ità (2)**(5:rep)**, -ismo (3)**(6:rep)**	**N: *2***	**N: *2***
5.	***-ino***	A rel	N -ismo (2)**(>10:rep)**, -ità (1)**(2:rep)**	**N: *2***	**N: *2***
6.	***-ismo***	N abstr	Ø	Ø	Ø
7.	***-ista***	N pers	Ø	Ø	Ø
8.	***-izzare***	V caus	N: -mento (1)**(4:rep)**, -zione (>100)**(>1000:rep)** (default for abstract nouns) -tore (21)**(>150:rep)** (derives objects) A: -bile (8)**(>100)** (default for adjectives) -(t)orio (1)**(10:rep)**	**N: *3*** **A: *2***	**N: *3*** **A: *2***
9.	***-oso***	A qualit	N -ità (>10), -ismo (3)**(4:rep)**, -ario(locat) (1)	**N: *3***	**N: *3***
10.	***-(t)orio***	A rel	Ø	Ø	Ø

Neither *derIvaTario* nor *la Repubblica* invalidates the cognitive approach we follow. Actually, *derIvaTario* differs from l*a Repubblica* with respect to existing SUFF2 suffixes in a single case, number 1, *-ese*$_A$ (Table 8). According to *derIvaTario*, *-ese*$_A$ can be followed by two suffixes for derivation of nouns, *-ita*$_N$ and *-ismo*$_N$, while in *la Repubblica* we could also find *-eria*$_N$. However, the combination *–ese+-eria*$_N$ derives only six types in *la Repubblica*, i.e., it should be rote-learned and does not influence our analysis of the derivational suffix combination in Italian.

Thus we come to the suffixes that according to both the *derIvaTario* and *la Repubblica* are never followed by other suffixes. In the literature on affix ordering, such suffixes are called closing. Closing suffixes have been reported in a number of languages: Szymanek (2000) is on closing morphemes (the term Szymanek uses) in English and Polish; Aronoff and Fuhrhop (2002) report a phenomenon that bans the further derivation in German and explain it in terms of closing suffixes; Manova (2008, 2010) provides evidence for closing suffixes in Bulgarian and Russian; Plungian and Sitchinava (2009) speak of closing suffixes in Russian; Melissaropoulou and Ralli (2010) acknowledge the existence of closing suffixes in Greek derivational morphology; and Manova and Winternitz (2011) discuss closing diminutive suffixes in Bulgarian and Polish. For closing suffixes always arises the question of what bans the further suffixation. As the equivalents of the Italian suffix *-ismo* are closing in all languages that have been investigated for closing suffixes so far (Manova 2015b), it is no surprise that *-ismo* is closing in Italian. However, why should the other two suffixes, *-ista*$_N$ and *-(t)orio*$_A$, be closing? Note that the parallel suffixes in Bulgarian allow further suffixation. In Bulgarian, all nouns that derive human beings serve as bases for derivation of collective and abstract nouns and relational or possessive adjectives; and all adjectives, except those derived by the suffix *-ski* (Manova 2008, 2015b) serve as bases for derivation of abstract nouns. Thus based on Bulgarian, we looked for specific words on the Internet that should contain combinations of *-ista*$_N$ and *-(t)orio*$_A$ with other suffixes, and such forms, though very rare, do exist. The examples of combinations that should not have existed according to *la Repubblica* corpus are listed in Table 9. This experiment with closing suffixes provides further evidence for a well-known fact, namely that even the largest electronic corpus does not contain all words in a language. Nevertheless, for research on affix ordering omissions of tokens in a corpus mean oversight of affix combinations. In our case, however, whether closing or non-closing, *-ista*$_N$ and *-(t)orio*$_A$ do not invalidate Manova's (2011b) Cognitive approach we follow in this paper, as the examples in Table 9 are extremely rare, i.e., they do not derive more than 10 types and should be rote-learned.

Table 9: Closing suffixes: *La Repubblica* versus the Internet

No in Table 8	**SUFF1**	**Lexical and semantic category of SUFF1**	**SUFF2 according to its lexical category**	**Examples**	**Translations**
6.	***-ismo***	N abstr	Ø	Ø	Ø
7.	***-ista***	N pers	A: *-ese* (internet)	*enigm-ist-ese*	language of puzzle games
10.	***-(t)orio***	A rel	N: *-età* (internet)	*sens-ori-età*	sensoriness

Thus, based on the comparison between the combinations of suffixes in *derIvatario* and *la Repubblica*, a small, specialized resource seems to be as reliable as a very large corpus. However, there is a big difference in the ways one works with a small and a huge resource. Working with a huge corpus requires much more time and effort than working with a small resource, irrespective whether a paper dictionary or an annotated electronic lexicon.

5. Conclusions

We have investigated the combinability of two sets of derivational suffixes: one from Bulgarian and another one from Italian and the goal was to establish what sources of data can be used for research on affix ordering in the languages under scrutiny. The so-called Cognitive approach to affix ordering (Manova 2011b) served as a theoretical framework. Both sources of data we used for Italian, the specialized annotated small lexicon *derIvaTario* and the huge *la Repubblica* corpus largely coincide regarding the suffixes that can follow a suffix. The corpus size does not seem to play a significant role for determining the affix combinations in a language and a corpus of relatively small size such as the Italian CoLFIS that contains a bit over 3 M tokens and on which the specialized lexicon *derIvaTario* is based appears large enough to be a reliable resource for research on affix ordering. For inflectional languages without specialized resources such as Bulgarian, a standard reverse dictionary of about 70,000 words is a good starting point for research on affix ordering.

References

ALLEN, Margaret. 1978. "Morphological Investigations." PhD diss., University of Connecticut.

ANDREJČIN, Ljubomir, ed. 1975. *Obraten rečnik na săvremennija bălgarski ezik*. Sofija: BAN. [Reverse dictionary of Modern Bulgarian.]

ARONOFF, Mark. 1976. *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

ARONOFF, Mark, FUHRHOP, Nanna. 2002. "Restricting Suffix Combinations in German and English: Closing Suffixes and the Monosuffix Constraint." *Natural Language and Linguistic Theory* 20: 451–490.

BARONI M., BERNARDINI S., COMASTRI F., PICCIONI L., ASTON G., MAZZOLENI M. 2004. Introducing the la Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In Proceedings of LREC 2004, pp. 1771–1774. Lisbon: ELDA.

BERTINETTO, Pier Marco, BURANI, Cristina, LAUDANNA, Alessandro, MARCONI, Lucia, RATTI, Daniela, ROLANDO, Claudia, THORNTON, Anna Maria. 2005. *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. Available online at: <http://linguistica.sns.it/CoLFIS/Home.htm>

Bălgarski Nacionalen Korpus / Bulgarian National Corpus. Available online at: <http://ibl.bas.bg/BGNC_bg.htm>

Bălgarski Referenten Korpus / Bulgarian Reference Corpus. Available online at: <http://www.webclark.org/help_en.html>

CROFT, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. New York: Oxford University Press.

derIvaTario. An annotated lexicon of Italian derivatives. Available online at: <http://derivatario.sns.it>

DRESSLER, Wolfgang. U., MAYERTHALER, Willi, PANAGL, Oswald, and WURZEL, Wolfgang U. 1987. *Leitmotifs in Natural Morphology*. Amsterdam: Benjamins.

DRESSLER, Wolfgang U. 1990. "The Cognitive Perspective of 'Naturalist' Linguistic Models." *Cognitive Linguistics* 1: 75–98.

DRESSLER, Wolfgang U. 2005. "Word-Formation in Natural Morphology." In *Handbook of Word-Formation*, edited by Pavol Štekauer and Rochelle Lieber , 267–284. Dordrecht: Springer.

FABB, Nigel. 1988. "English Suffixation is Constrained only by Selectional Restrictions." *Natural Language and Linguistic Theory* 6: 527–539.

FILLMORE, Charles. 1982. "Frame semantics." In *Linguistics in the Morning Calm*. Seoul, Hanshin Publishing Co., 111-137.

GEERAERTS, D. 2006. *Cognitive Linguistics: Basic Readings*. Berlin: Mouton de Gruyter.

GIEGERICH, H. 1999. *Lexical Strata in English: Morphological Causes, Phonological Effects*. Cambridge: Cambridge University Press.

HAY, Jennifer. 2001. "Lexical Frequency in Morphology: Is Everything Relative?" *Linguistics* 39: 1041–1070.

HAY, Jennifer. 2002. "From Speech Perception to Morphology: Affix-ordering Revisited." *Language* 78: 527–555.

HAY, Jennifer. 2003. *Causes and Consequences of Word Structure*. London: Routledge.

HAY, Jennifer, PLAG, Ingo. 2004. "What Constrains Possible Suffix Combinations? On the Interaction of Grammatical and Processing Restrictions in Derivational Morphology." *Natural Language and Linguistic Theory* 22: 565–596.

JACKENDOFF, Ray. 1990. *Semantic Structures*. Cambridge, MA: MIT Press.

KIPARSKY, Paul. 1982. "Lexical Morphology and Phonology." In *Linguistics in the Morning Calm: Selected Papers from SICOL-1981 (vol. 1)*, edited by I.-S. Yang, 3–91. Seoul: Hanshin.

LANGACKER, R. (1987). *Foundations of Cognitive Grammar, Volume I, Theoretical Prerequisites.* Stanford University Press.

LAUDANNA, Alessandro, THORNTON, Anna Maria, BROWN, Giorgina., BURANI, Cristina, MARCONI, Lucia. 1995. Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente, in S. Bolasco, L. Lebart e A. Salem (ed.) *III Giornate internazionali di Analisi Statistica dei Dati Testuali. Volume I*, (Roma: Cisu), 103-109.

MANOVA, Stela. 2008. „Closing suffixes and the structure of the Slavic word: Movierung." *Wiener Slawistisches Jahrbuch* 54: 91–104.

MANOVA, Stela. 2010. "Suffix Combinations in Bulgarian: Parsability and Hierarchy-based Ordering." *Morphology* 20: 267–296.

MANOVA, Stela, ARONOFF, Mark. 2010. "Modeling Affix Order." *Morphology* 20: 109–131.

MANOVA, S. 2011a. *Understanding Morphological Rules*. Dordrecht: Springer.

MANOVA, Stela. 2011b. "A Cognitive Approach to SUFF1-SUFF2 Combinations: A Tribute to Carl Friedrich Gauss." *Word Structure* 4: 272–300.

MANOVA, Stela, WINTERNITZ, Kimberly. 2011. "Suffix Order in Double and Multiple Diminutives: With Data from Polish and Bulgarian." *Studies in Polish Linguistics* 6: 115–138.

MANOVA, Stela. 2014. "Affixation." *Oxford Bibliographies in Linguistics*, edited by Mark Aronoff. New York: Oxford University Press. DOI: 10.1093/OBO/9780199772810-0183. Available online at: http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0183.xml.

MANOVA, Stela. 2015a. "Affix order and the Structure of the Slavic Word." In *Affix Ordering Across Languages and Frameworks*, edited by Stela Manova, 205–229. New York: Oxford University Press. DOI:10.1093/acprof:oso/9780190210434.003.0009. Oxford Scholarship Online: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780190210434.001.0001/acprof-9780190210434-chapter-9

MANOVA, Stela. 2015b. "Closing Suffixes." In *Word Formation: An International Handbook of the Languages of Europe*, Vol. 2, edited by Peter Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer. Handbooks of Linguistics and Communication Science (HSK) 40/2. Berlin: De Gruyter Mouton, 956-971.

MERLINI BARBARESI, Lavinia. 2012. "Combinatorial patterns among Italian evaluative affixes." *SKASE Journal of Theoretical Linguistics* 9(1): 2-14.

MESTRES-MISSÉ, A., RODRIGUEZ-FORNELLS, A., MÜNTE, T. F. 2010. "Neural Differences in the Mapping of Verb and Noun Concepts onto Novel Words." *Neuromage* 49: 2826–2835.

MELISSAROPOULOU, Dimitra, RALLI, A. 2010. "Greek derivational structures: restrictions and constraints." Morphology 20 (2): 343–357.

MOHANAN, K. P. 1986. *The Theory of Lexical Phonology*. Dordrecht: Reidel.

MURDAROV, V., ALEKSANDROVA, T., DIMITROVA, M., STANČEVA, R., ČARALOZOVA, K., TOMOV, M. (2011). *Obraten rechnik and bălgarskija ezik* / A reverse dictionary of Bulgarian. Sofija: Iztok-Zapad.

la Repubblica corpus. Available online at:
http://dev.sslmit.unibo.it/corpora/corpus.php?path&name=Repubblica

PLAG, Ingo. 1996. "Selectional Restrictions in English Suffixation Revisited. A Reply to Fabb (1988)." *Linguistics* 34: 769–798.

PLAG, Ingo. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Berlin and New York: Mouton De Gruyter.

PLAG, Ingo. 2002. "The Role of Selectional Restrictions, Phonotactics and Parsing in Constraining Suffix Ordering in English." In *Yearbook of Morphology 2001*, edited by Geert Booij and Jaap van Marle, 285–314. Dordrecht: Kluwer.

PLAG, Ingo, BAAYEN, Harald. 2009. "Suffix Ordering and Morphological Processing." *Language* 85: 109–152.

Reverse dictionary of modern Bulgarian Andrejčin (ed.) 1975.

SELKIRK, Elisabeth. 1982. *The Syntax of Words*. Cambridge, MA: MIT Press.

SIEGEL, Dorothy. 1974. *Topics in English Morphology*. Cambridge, MA: MIT Press.

SITCHINAVA, Dmitri, PLUNGIAN, Vladimir. 2009. "Closing suffixation patterns in Russian, with special reference to the Russian National Corpus. Paper presented at the *2nd Vienna Workshop on Affix Order: Affix Order in Slavic and Languages with Similar Morphology*, Vienna, June 5–6, 2009.

SKALIČKA, Vladimir. 1979. *Typologische Studien*, edited by Peter Hartmann. Wiesbaden:

Braunschweig.
SZYMANEK, Bogdan. 2000. "On morphotactics: Closing morphemes in English." In: Bożena Rozwadowska (ed.), *PASE Papers in Language Studies*, 311–320. Wrocław: Aksel.
SZYMANEK, Bogdan, DERKACH, Tetyana. 2005. "Constraints on the Derivation of Double Dminutives in Polish and Ukrainian." *Studies in Polish Linguistics* 2: 93–112.
ŠTEKAUER, Pavel. 1998. *An Onomasiological Theory of English Word-Formation*. Amsterdam: Benjamins.
TALAMO, Luigi. 2015. "Suffix Combinations in Italian: Selectional Restrictions and Processing Constraints." In *Affix Ordering Across Languages and Frameworkd*, edited by Stela Manova, 175–204. New York: Oxford University Press. DOI:10.1093/acprof:oso/9780190210434.003.0008. Oxford Scholarship Online: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780190210434.001.0001/acprof-9780190210434-chapter-8.
TALAMO, Luigi, CELATA, Chiara. 2011. "Toward a morphological analysis of the Italian lexicon: developing tools for a corpus-based approach." *Quaderni del Laboratorio di Linguistica*, 1(10)
TALAMO, Luigi, CELATA, Chiara, BERTINETTO, Pier Marco, in press. "DerIvaTario: An annotated lexicon of Italian derivatives." *Word Structure*.
TAYLOR, John R. 2002. *Cognitive Grammar*. New York: Oxford University Press.

[i] The first author was supported by a ESF grant, NetWordS-09-RNP-089 / individual grant 5566, the second author was funded by Scuola Normale Superiore (SNS), Pisa. The research reported herein was carried out during a research stay of the first autor at SNS and a portion of this work was presented at the linguistic seminar there. We thank the SNS students for their insightful comments and Prof. Pier Marco Bertinetto for his support to our research and the outstanding conditions we had in Pisa.

[ii] Abbreviations used in the text: abstr – abstract, A – adjective, caus – causative, durat – durative, IMPFV – imperfective, inchoat – inchoative, infl – inflection, N – noun, qualit – qualitative, rel – relational, SUFF – suffix, V – verb.

[iii] In Bulgarian, 'pharmacist's' is a possessive adjective.

[iv] Luigi Talamo wishes to thank Eros Zanchetta (University of Bologna) for granting him the access to La Repubblica corpus.

Stela Manova
University of Vienna
Department of Philosophy
Universitätsstraße 7
1010 Wien
e-mail: stela.manova@univie.ac.at

Luigi Talamo
University of Bergamo
Piazzetta Verzeri 1
24129 Bergamo
e-mail: luigi.talamo@unibg.it