# Investigating English Affixes and their Productivity with Princeton WordNet

**Verginica Barbu Mititelu**
Romanian Academy Research Institute for Artificial Intelligence
Bucharest, Romania
vergi@racai.ro

## Abstract

Such a rich language resource like Princeton WordNet, containing linguistic information of different types (semantic, lexical, syntactic, derivational, dialectal, etc.), is a thesaurus which is worth both being used in various language-enabled applications and being explored in order to study a language. In this paper we show how we used Princeton WordNet version 3.0 to study the English affixes. We extracted pairs of base-derived words and identified the affixes by means of which the derived words were created from their bases. We distinguished among four types of derivation depending on the type of overlapping between the senses of the base word and those of the derived word that are linked by derivational relations in Princeton WordNet. We studied the behaviour of affixes with respect to these derivation types. Drawing on these data, we inferred about their productivity.

## 1 Introduction

Affixes productivity, i.e. their use to create new words, can be studied on a corpus or on lists of words, in particular on dictionaries. Working with a corpus has several advantages over working with a dictionary: words are seen "in action" (i.e. one can see in what contexts they are used, in what forms, with what frequency, etc.); one can find words that are not recorded in dictionaries, either because they are brand new creations or because they are obtained in a (highly) regular way by a very productive word formation rule; frequencies can be counted for either types or tokens. However, we chose Princeton WordNet (PWN) (Fellbaum, 1998) version 3.0 for studying the productivity of English affixes. We wanted to test whether affixes productivity is influenced by the number of senses of the base form and of the derived word that are semantically unrelated. PWN has several characteristics that make it appropriate for our investigation. It contains quite a large number of words (155,287 lemmas) organized according to their senses (thus reaching 206,941 word-sense pairs)[1]. PWN also displays lexical density: "all" senses of a word are included; this is a great asset for our experiment, which is run at the word sense level.

The hypothesis of our study is that the meaning of the derived word is compositional, being a function of the meaning of the base word and of the affix(es) contained (other authors (Plag, 1999) formulate this as a function of the meaning of the rule and of the base). Whenever no semantic resemblance can be found between the two (in other words, derived words have an idiomatic meaning rather than a compositional one – see Bauer et al. (2013)) we do not consider them a derived-base pair of words. Nevertheless, we presume that the original meaning(s) of the derived words is/are (a) compositional one(s), whereas the idiomatic one(s) is/are the result of a semantic evolution in independence of the semantic evolution of its base word.

## 2 Related work

There are two lines of research interesting as background for our experiment: one has to do with the study of affixes productivity, and the other concerns the derivational morphology studies in connection to PWN or with other wordnets, each of them detailed in a separate subsection in what follows.

[1]The data are taken from http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html.

### 2.1 Affixes productivity

An affix is a morpheme that is attached to a word in order to create a new word, process known as derivation. Not all affixes in a language are productive to the same extent: some are more productive than others, while others may show no productivity at all; still others may cease being productive for some time and may get "reactivated" afterwards. Productivity is studied in synchrony: from one period to another one can notice differences in the productivity of the same affix, as said before.

Word formation processes, derivation included, are never totally unrestricted (Plag, 1999). Several factors have been discussed with respect to their influence on affixes productivity. On the one hand, there are both linguistic and non-linguistic ones; on the other hand, they show the interdependence of the various subsystems of the language (Aronoff, 1976). These factors are: morphological restrictions on the base word, semantic coherence (Aronoff, 1976), paradigmatic factors (van Marle, 1985), lexical government, lexical listing, phonological factors (Aronoff, 1976; Baayen, 1992), phonotactics (Hay and Baayen, 2003), etymology of the base word (Bauer et al., 2013), parsing (i.e. decomposition in perception) (Hay and Baayen, 2002), type and token frequency (Baayen, 1992), contextual appropriateness (Burgschmidt, 1977), socio-economic status of the language user and his/her attitude towards linguistic phenomena (Baayen, 1992), "fashion" (Plag, 1999).

### 2.2 Derivational morphology and wordnets

Several wordnets (American (Fellbaum et al., 2009), Czech (Pala and Hlaváčková, 2007), Bugarian (Koeva, 2008; Dimitrova et al., 2014; Koeva et al., 2016), Romanian (Barbu Mititelu, 2012), among others) have gone beyond their original structure and included, between pairs of literals, new relations, derivational in nature: the connected literals are the base and the derived words, of course considered with their respective meaning (from the synset to which they belong). Such relations reflect both the formal connection between the two literals (i.e. one is created from the other by means of derivation, that is by adding an affix to it) and the semantic connection: the derived literal has a compositional meaning, in which one can recognize the meaning of the base word and the contribution of the affix. Either manually or automatically, the pairs are identified and labeled using various sets of relation names. Such relations are identified for certain parts of speech (as is the case in Bulgarian (Koeva et al., 2016), Croatian (Koeva, 2008) or American wordnets, among others) or all of them (e.g., Polish (Piasecki et al., 2012) and Romanian (Barbu Mititelu, 2012), among others) and are labeled differently from one wordnet to the other, although some overlaps exist.

In the projects enriching wordnets with such relations there has been interest in making these resources richer and more useful for various applications (Barbu Mititelu, 2013).

## 3 The experiment

In this section we present an experiment in which we extracted the pairs of base - derived word from PWN and assigned them to a different class according to the way their senses are related by a derivational relation.

### 3.1 Aim

The hypothesis we wanted to test here and that had not been touched upon in any previous study that we are aware of is whether the number of senses the base word and the derived word, the proportion of them being interlinked and/or the semantic evolution of the derived word independently from the base are factors that could influence affixes productivity.

### 3.2 Data preparation

Among the relations marked in PWN v. 3.0 there are several that link pairs of derivationally related words: `derivat` (linking nouns to their noun, verb or adjective roots, verbs to their noun or adjective roots, adjectives to their noun, verb or adverb roots, and adverbs to their adjective roots), `derived_from` (linking adverbs derived from adjectives), `pertainym` (linking adjectives to their noun roots). We extracted all pairs of words linked by the first two relations mentioned. The last one (`pertainym`) was disregarded because it usually doubles the relation `derivat`, i.e. it links words that are usually also linked by the `derivat` relation, as in the following example: the adjective *academic* in its first sense establishes two relations with the noun *academia*: one is `derivat` and the other one is `pertainym`.

We extracted 77,939 pairs of words (base -

derived word) between which there is either a `derivat` or a `derived_from` relation. However, some of them are duplicates: for example, the adjective *scarce* is related to the nouns *scarcity* and *scarceness* by means of the relation `derivat`; in their turn, both nouns are linked to the adjective *scarce* by means of the relation `derivat`. Thus, we eliminated duplicates in the data and were left with 40,632 pairs. We added 73 pairs which involved participles linked to their base verbs by means of the relation `participle`: for example, *avenged* (marked as adjective) is linked to the verb *avenge* by means of the relation `participle`.

Further cleaning of the data was done in order to eliminate dialectal duplicates: words belonging to the same synsets and that differ in the spelling with *-ise* or *-ize*, on the one hand, and words containing the *-ou-* or the *-o-* sequence, on the other hand: examples: *equalise - equalize*; *discolouration - discoloration*. Only one of the pairs was kept, in each case. The former type of duplicates occurred 81 times in the data, while the latter occurred 306 times.

Thus, the list we focused on for annotation contained 40,318 pairs of base - derived words, including all parts of speech in PWN.

### 3.3 Data annotation

For all these pairs we automatically extracted the affix(es). The base and the derived words were compared as strings of letters and the difference found between them was checked against a list of English affixes containing 26 prefixes and 54 suffixes. In case the string was found in that list, it was considered an affix and marked as such in the annotation. Otherwise, manual intervention (by one linguist) was necessary for identifying the affix(es) or their combination in case of parasynthetic derivation (i.e. by means of both a prefix and a suffix) or successive derivation. During the manual inspection of the pairs we also identified pairs that are in no derivational relation at all: *inappropriate* and *wrongness*, *immunology* and *allogeneic*, etc. They were eliminated from the data. Another situation is that of words like *skepticism - skeptical*: they are both created from the same root, *skeptic*, each with a different suffix: *-ism* and, respectively, *-al*, so they are not derived one from the other. Such pairs were also disregarded, just like cases of a similar type: *atheism - atheistic*, where one can recognize the Greek elements *a-* and *theos*, but the former is borrowed from French (where the word was obtained by adding the suffix *-isme* to the Greek elements) and the latter is derived in English by adding the suffix *-ic* to the French borrowing *athéiste* (itself derived by adding the suffix *-iste* to the Greek *atheos*). Thus, the total number of annotated pairs was 30,018.

For all these pairs we identified the affix, we extracted from PWN the number of senses each of the literals in the pairs has and the number of derivational relations established between the two literals. Afterwards, we counted:

- the number of senses with which the base word participates in the derivational links with the derived words
- their percent in the total number of senses of the base word
- the number of senses the derived word participates in the derivational links with the base
- their percent in the total number of senses of the derived word.

It is important to note that the numbers representing the number of derivational relations established between the two literals, the number of senses with which the base word participates in derivational links with the derived word, and the number of senses with which the derived word participates in the derivational links with the base need not be identical. Let us consider the following pair: *buzz - buzzer*. The verb base word has the following senses:

- *buzz:1* - make a buzzing sound
- *buzz:2* - fly low
- *buzz:3* - be noisy with activity
- *buzz:4* - call with a buzzer

The derived noun has the following senses:

- *buzzer:1* - a push button at an outer door that gives a ringing or buzzing signal when pushed
- *buzzer:2* - a signaling device that makes a buzzing sound

The four derivational relations established between the two words are as follows:

- *buzz:1 - buzzer:1*
- *buzz:1 - buzzer:2*
- *buzz:4 - buzzer:1*
- *buzz:4 - buzzer:2*

There are four derivational relations between the two words, but, whereas all senses of the derived word enter these relations, only two out of the four senses of the base participates to them.

Another step in the annotation was the automatic identification of the derivation type, as we will explain below. We automatically counted the number of senses specific to the base word, i.e. not establishing links with the derived word, the number of senses specific to the derived word, and the ratio between the senses specific to the derived word and those specific to the base word.

Four types of derivation were identified as types of sets intersection. Whenever **all** senses of the derived word are linked to **some** of the senses of the base word, we mark the pair as being of the **R** type: see the pair *buzz - buzzer* above. When **some** senses of the derived word are derivationally linked to **all** of the senses of the base word, we mark the pair as being of the **D** type: see *restitute - restitution*: the base verb has the following senses:

- *restitute:1* - give or bring back
- *restitute:2* - restore to a previous or better condition

The derived noun has the following senses:

- *restitution:1* - a sum of money paid in compensation for loss or injury
- *restitution:2* - the act of restoring something to its original state
- *restitution:3* - getting something back again

The derivational relations established between the two words are as follows:

- *restitute:2 - restitution:2*
- *restitute:1 - restitution:3*

Both senses of the base are linked to some of the senses of the derived word.

In case of identical sets, which means that there is no sense of the base word that is not derivationally linked to any of the senses of the derived word and vice versa, there is no sense of the derived word that is not linked to any of the senses of the base word, we mark the pair as being of the **RD** type: see the pair *explore - exploration*: the base verb has the following senses:

- *explore:1* - inquire into
- *explore:2* - travel to or penetrate into
- *explore:3* - examine minutely
- *explore:4* - examine (organs) for diagnostic purposes

The derived noun has the following senses:

- *exploration:1* - to travel for the purpose of discovery
- *exploration:2* - a careful systematic search
- *exploration:3* - a systematic consideration

The derivational relations established between the two words are as follows:

- *explore:1 - exploration:3*
- *explore:2 - exploration:1*
- *explore:2 - exploration:3*
- *explore:3 - exploration:2*
- *explore:3 - exploration:3*
- *explore:4 - exploration:2*

All senses of both words are involved in these six derivational links between them.

When at least one sense of the derived word is linked to at least one sense of the base word, and there is at least one sense of the derived word not linked to any sense of the base word and at least one sense of the base word not linked to any sense of the derived word, we mark the pair as being of the **I** type: see *perform - performance*: the base verb has the following senses:

- *perform:1* - carry out or perform an action
- *perform:2* - perform a function
- *perform:3* - give a performance (of something)
- *perform:4* - get (something) done

The derived noun has the following senses:

- *performance:1* - a dramatic or musical entertainment
- *performance:2* - the act of presenting a play or a piece of music or other entertainment
- *performance:3* - the act of performing; of doing something successfully; using knowledge as distinguished from merely possessing it
- *performance:4* - any recognized accomplishment
- *performance:5* - process or manner of functioning or operating

There are only two derivational relations established between the two words, involving only a couple of their senses:

- *perform:1 - performance:3*
- *perform:3 - performance:1*

All the other senses of the two words remain derivationally unrelated.

For each affix (or combination of affixes) we calculated the frequency of the different types of derivation (R, D, RD, I) to which it participates in PWN (see subsection 4.2 below for the interpretation of these data).

## 4 Results and their linguistic significance

There are several results of this undertaking. One of them is the list of pairs extracted from PWN and enriched with information as described above. We discuss the others in the subsections below.

### 4.1 Derivation types

The total number of occurrences of the derivation types is 30,018. The most frequent one is the RD type - 12,792 occurrences. The second most frequent one is the R type (11,043 occurrences). They are followed, at long distance, by type I (4,267 occurrences) and type D (1,916 occurrences).

The highest frequency of the RD type shows that most of the derived words share the meanings of their base. However, there is also a large number of cases when the derived word is "semantically less rich" than its base word - see the high number of occurrences of type R.

Much less frequent (4,267) is the case of pairs in which the two words have both meanings in common (type I), and an independent semantic evolution. This is the case of pairs such as *dust - duster*. The former has the following meanings:

- *dust:1* - remove the dust from
- *dust:2* - rub the dust over a surface so as to blur the outlines of a shape
- *dust:3* - cover with a light dusting of a substance
- *dust:4* - distribute loosely

The latter has the meanings:

- *duster:1* - a windstorm that lifts up clouds of dust or sand
- *duster:2* - a loose coverall (coat or frock) reaching down to the ankles
- *duster:3* - a piece of cloth used for dusting
- *duster:4* - a pitch thrown deliberately close to the batter

Only *dust:1* is derivationally related to *duster:3*. The other meanings remain semantically distant.

We should note that types R and RD may contain false positives examples, because in wordnets there is no distinction between polysemous words and homographs of the same part of speech: they are both recorded as different senses of the same literal.

The least frequent (1,916) is the case of derived words that develop new meanings (after derivation) (type D): consider the adjective *amphibious* derived from *amphibia*. Besides the meaning "relating to or characteristic of animals of the class Amphibia", which clearly links it to the base (having the meaning "the class of vertebrates that live on land but breed in water; frogs; toads; newts; salamanders; caecilians"), the derived word has developed another meaning ("operating or living on land and in water"), which applies to various semantic types of nouns, as the examples in PWN show: "amphibious vehicles"; "amphibious operations"; "amphibious troops"; "frogs are amphibious animals", in complete independence from the base.

In terms of affixes productivity, only types D and I are interesting: we can think of the new

meanings of the derived words in PWN as hapax phenomena (i.e., the words occurring only once in PWN) in a corpus. Consequently, following (Baayen, 1992), who proved that the number of hapax legomena instances of words derived with a certain affix in a corpus is suggestive of that affix productivity, we can consider affixes involved in these two types of derivation to be productive ones (see the next subsection).

### 4.2 Affixes and types of derivation

Having annotated the type of derivation pertinent to each pair, we can test if affixes manifest any affinity with these derivation types.

A first remark on the data is that affixes rarely tend to belong to only one derivation type. We looked at the ten most frequent ones in our data. They are:

- *-ness* - 3,730 occurrences;
- *-er* - 3,100 occurrences;
- *-ly* - 2,953 occurrences;
- *-ion* - 2,469 occurrences;
- *-ing* - 2,102 occurrences;
- *-ation* - 1,546 occurrences;
- *-ic* - 1,290 occurrences;
- *-ity* - 1,186 occurrences;
- *-al* - 1,011 occurrences;
- *-ist* - 805 occurrences.

Their distribution according to the four types of derivation is rendered in Figure 1 below. All these affixes participates in all four types of derivation, even if to a different extent. We can note that the RD type is predominant for most affixes, except for *-ing*, *-ly* and *-er*, which tend to participate in derivations of type R.

Type R of derivation tends to be realized by the affixes *-ly*, *-er*, *-ness*, *-ing*, as obvious in Figure 2. Type RD is realized by the affix *-ness* to the highest extent. Type D is more frequently realized by the affix *-ion*, almost three times more often than the next frequent affix for this derivation type, namely *-ation*. Type I is realized mostly by the suffixes *-er* and *-ion* and, to a lesser and comparable extent, by the other suffixes in the top 10 most frequent ones in our data.

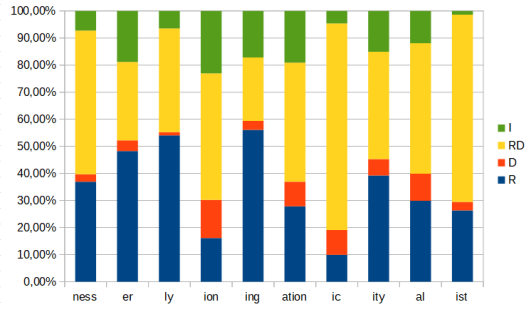


Figure 1: The 10 most frequent affixes and the frequency of the types of derivation to which they participate.

Little correlation can be noted between the affixes realizing the D and I types of derivation. Besides the prevalence of the suffix *-ion* with both types, nothing else strikes us when comparing the two.

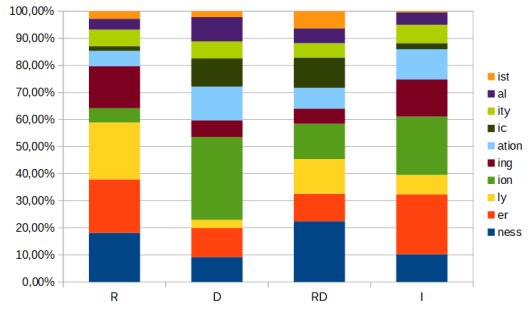


Figure 2: The four types of derivation and the affixes involved.

### 4.3 Affixes productivity

We compared the data we obtained with the statistical data about affixes provided by Hay and Baayen (2002). They report on a corpus-based research: their calculations "are based on a set of words extracted from the CELEX Lexical Database (Baayen et al., 1995)". We noted a correlation of their results with the PWN-based data obtained by us.

Firstly, the frequency of affixes is similar in the two experiments: looking only at the most frequent ones, the following affixes occur on both lists: *-er*, *-ly*, *-y*, *-ness*, *-al*, *-ic*, *-ity*, *-able*. Hay and Baayen (2002) also report a high frequency of the suffixes *-like* and *-less*. The former has only one occurrence in our data, whereas the latter is completely absent: words derived with *-less* (such

as *harmless*, *speechless*, etc.) are not derivationally related in PWN to their respective bases.

Secondly, comparing the number of hapax legomena for individual affixes in the corpus-based experiment with the sum of the frequency of D type and I type derivations for the same affixes in the PWN-based experiment, we also notice similarities between data: the most productive affixes, from both perspectives, are: *-er*, *-y*, *-ly*, *-ness*. Other very productivee ones are: *-or*, *-able*, *-an*. They all display a high number of hapaxes in the corpus and, respectively, high number of total occurrences in derivations of types D and I.

## 5 Conclusions and future work

A mature resource, PWN can be used, besides in language-enabled applications, in linguistic studies of various types. Our experiment is grounded in the assumption that derivation is a relation between word senses rather than between words as sets of meanings. This relation manifests in a formal and semantic way: formally, one word (the derived one) in the relation is obtained from the other (the base word) (usually) by adding some linguistic material (an affix); semantically, the meaning of the derived word is compositionally obtained from the meaning of the base word and of the affix(es) it contains. PWN follows this assumption and, thus, offers the perfect environment for testing the hypothesis that affixes that are involved in deriving words that develop meanings independently from their base word are morphologically productive ones. As shown above, this seems to be the case.

We have also presented here, based on the data extracted from PWN and annotated, information about affixes frequency in general and, in particular, their frequency depending on four types of derivation defined ad hoc, thus their tendencies to participate in one type or another of derivation.

However, as obvious from the discussion in this paper, the degree of coverage and of correctness of the derivational links in PWN varies from one affix to the other. It is straightforward that this fact has an impact on our research. Nevertheless, we could not evaluate it for this presentation of results.

As further work, we could also check if PWN granularity, already proved to be too fine, is reflected in the way derivation is marked in the network: for this, we would look, for each derived literal, at the number of derivational links each of its senses establishes with its base word.

Other aspects of affixes study that can be extracted from further processing the data we now have are: affixes capacity of allowing for the inheritance by the derived word of the meaning(s) of the base word (calculated as the percent of senses of the base word that are linked to the derived word), their capacity of allowing sense evolution (calculated as the percent of senses specific to the derived word) and the ratio of the derived word specific senses and of the base word specific senses.

The semantic types of the base words to which one affix can attach is another line of research possible to be explored with our data.

Our experiment could be repeated for another language for which there is quite a large wordnet, in whose development the implementation of as many senses of a word as possible was an objective.

## Acknowledgments

I would like to express my gratitude to Prof. Harald R. Baayen for the insightful discussions we had about this topic.

## References

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. Cambridge, MA, London, England: MIT Press.

Harald Baayen. 1992. *Quantitative aspects of morphological productivity*. In G. E. Booij and J. van Marle (eds.). *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers:109–149.

R.H. Baayen, R. Piepenbrock, and L. Gulikens. 1995. *The CELEX lexical database (release 2) cd-rom*. Philadelphia, PA: Linguistic Data Consortium, Univesity of Pennsylvania.

Verginica Barbu Mititelu. 2012. *Adding morpho-semantic relations to the Romanian Wordnet*. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*:2596–2601.

Verginica Barbu Mititelu. 2013. *Increasing the Effectiveness of the Romanian Wordnet in NLP Applications*. *Computer Science Journal of Moldova*, vol. 21, no. 3(63):320-331.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press.

Ernst Burgschmidt. 1977. *Strukturierung, Norm und Productivität in der Wortbildung*. In H. E. Brekle

and D. Kastovsky (eds.). *Perspektiven der Wortbildungforschung*. Bonn: Bouvier Verlag.

Tsvetana Dimitrova, Ekaterina Tarpomanova, and Borislav Rizov. 2014. *Coping with derivation in the Bulgarian WordNet*. In *Proceedings of the Seventh Global WordNet Conference (GWC 2014)*:109–117.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. *Putting semantics into WordNets "morphosemantic" links*. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technology*. Springer Lecture Notes in Informatics, volume 5603:350-358.

Jeniffer Hay and Harald Baayen. 2002. *Parsing and Productivity*. In G. E. Booij and J. van Marle (eds.). *Yearbook of Morphology*. Dordrecht: Kluwer Academic Publishers:203–235.

Jennifer Hay and Harald Baayen. 2003. Phonotactics, Parsing and Productivity. *Italian Journal of Linguistics*, 1:99–130.

Svetla Koeva. 2008. Derivational and morphosemantic relations in Bulgarian Wordnet. *Intelligent Information Systems*, 359–368.

Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova, Tsvetana Dimitrova, and Maria Todorova. 2016. *Automatic Prediction of Morphosemantic Relations*. In *Proceedings of the Eighth Global WordNet Conference (GWC 2016)*:168–176.

Karel Pala and Dana Hlaváčková. 2007. *Derivational relations in Czech WordNet*. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*:75–81.

Maciej Piasecki, Radosław Ramocki, and Marek Maziarz. 2012. *Recognition of Polish Derivational Relations Based on Supervised Learning Scheme*. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*:916–922.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: De Gruyter.

Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, Vol 0, No 1:111–142.

Jaan van Marle. 1985. *On the Paradigmatic Dimensions of Morphological Creativity*. Dordrecht: Foris.