Investigating English Affixes and their Productivity with Princeton WordNet

Verginica Barbu Mititelu

Romanian Academy
Research Institute for Artificial Intelligence

Introduction

- PWN – mature resource
- Assumption: derivation is a relation between word senses; manifest at two levels:
  - Form: the derived word is obtained from the base word (usually) by adding some linguistic material (an affix)
  - Meaning: the meaning of the derived word is compositionally obtained from the meaning of the base word and of the affix(es) it contains
- Question: is there any correlation between the affixes productivity and the types of derivation they are involved in?

Factors that influence affixes productivity

- morphological restrictions on the base word
- semantic coherence
- paradigmatic factors
- lexical government
- lexical listing
- phonological factors
- phonotactics
- etymology of the base word
- parsing (i.e. decomposition in perception)
- type and token frequency
- contextual appropriateness
- socio-economic status of the language user and his/her attitude towards linguistic phenomena
- ‘’fashion’’

- What about:
  - the number of senses of the base word and of the derived word
  - the proportion of them being interlinked
  - the semantic evolution of the derived word independently from the base

The data

- Princeton WordNet v. 3.0
- Pairs of literals linked by one of the relations:
  - derivat
  - derived_from
  - participle

    =>78,012 pairs

Data cleaning

- Keep one of the duplicates only:

  ex.: *scarce* – *scarcity*; *scarcity* – *scarce*

- Eliminate dialectal duplicates:

  ex.: -ise/-ize (*equalise*/*equalize*); -ou-/-o- (*discolouration*/*discoloration*)

- Left pairs: 40,318

Data annotation

- Automatic identification of affixes (26 prefixes, 54 suffixes) + manual intervention
- Treatment of special cases:
  - Analysable borrowings are annotated;
  - Successive derivation is marked as one step:

    e.g. *argue* - *argumentation*
- Disregarded cases:
  - No common root: *innappropriate* – *wrongness*;
  - Words derived from the same root with different affixes: *skepticism* – *skeptical*
- Annotated pairs: 30,018

Expanding the data

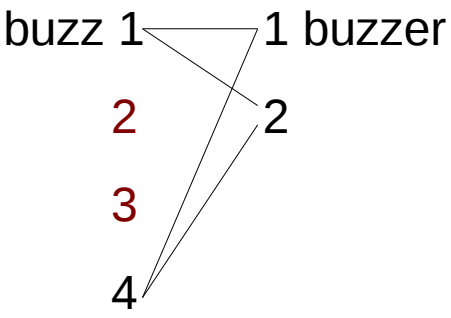
- For all pairs, extract from PWN 3.0:
  - the number of senses with which the base word participates in the derivational links with the derived word
  - their percent in the total number of senses of the base word
  - the number of senses the derived word participates in the derivational links with the base
  - their percent in the total number of senses of the derived word

Types of derivation

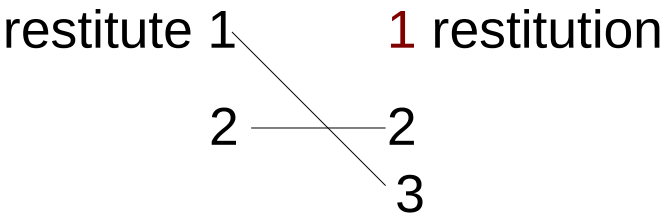
- R: **all** senses of the derived word are linked to **some** of the senses of the base word
- D: **some** senses of the derived word are derivationally linked to **all** of the senses of the base word
- RD: there is no sense of the base word that is not derivationally linked to any of the senses of 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
- I: 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

Types of derivation

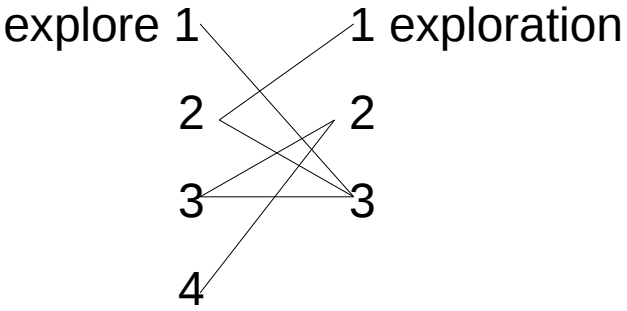
- R



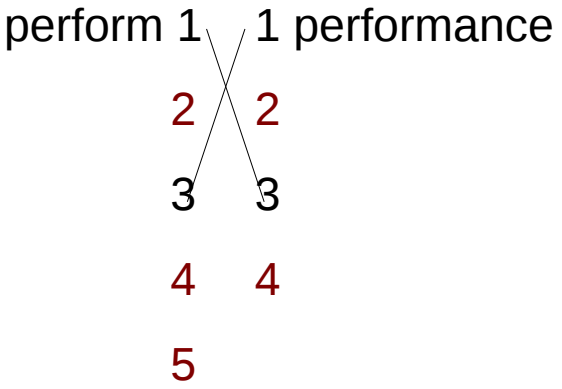
- D



- RD



- I



Results and their linguistic significance

- 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 the annotated pairs.

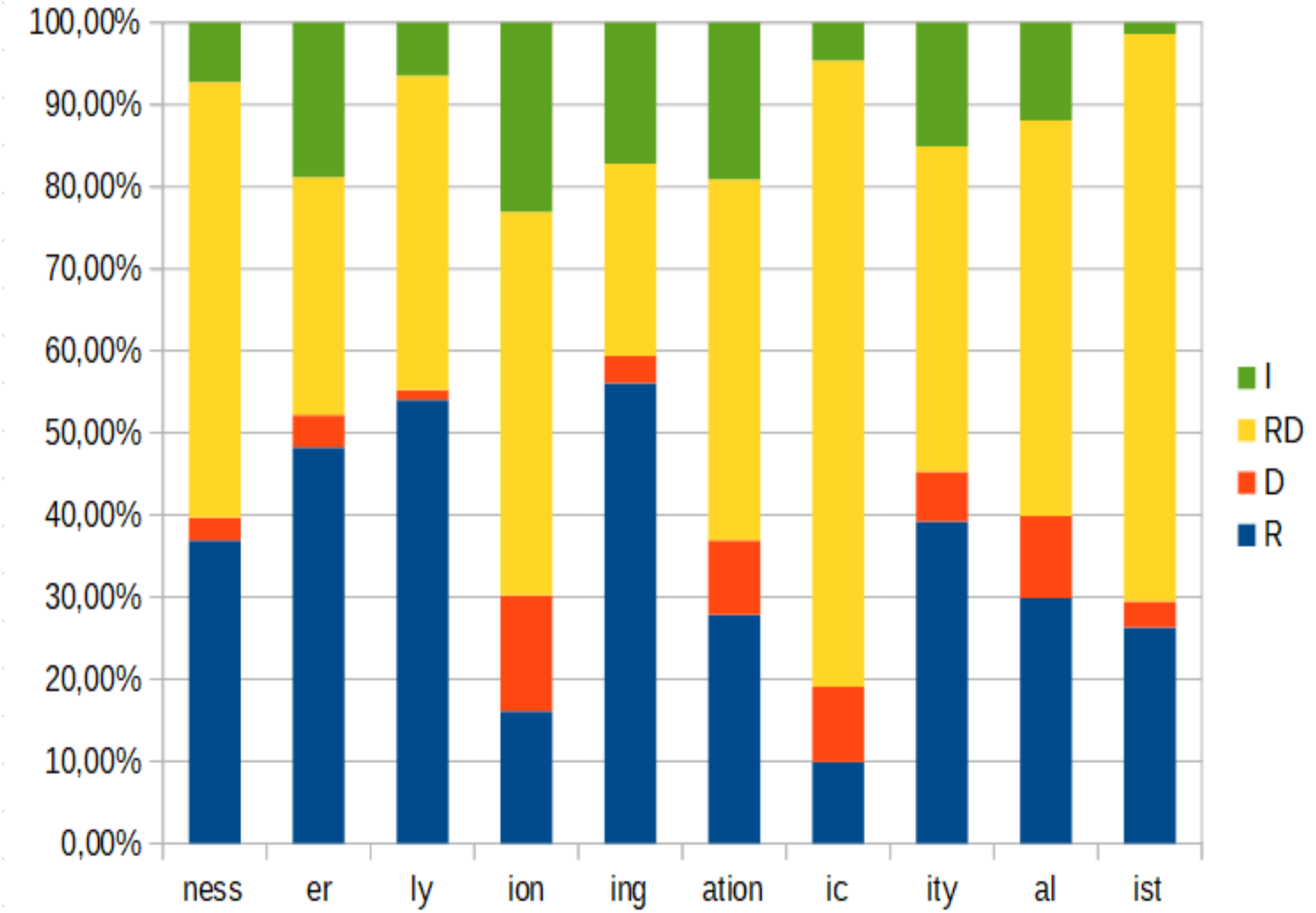
Derivation types

- Total number: 30,018
- Types distribution:
  - RD: 12,792
  - R: 11,043
  - I: 4,267
  - D: 1,916

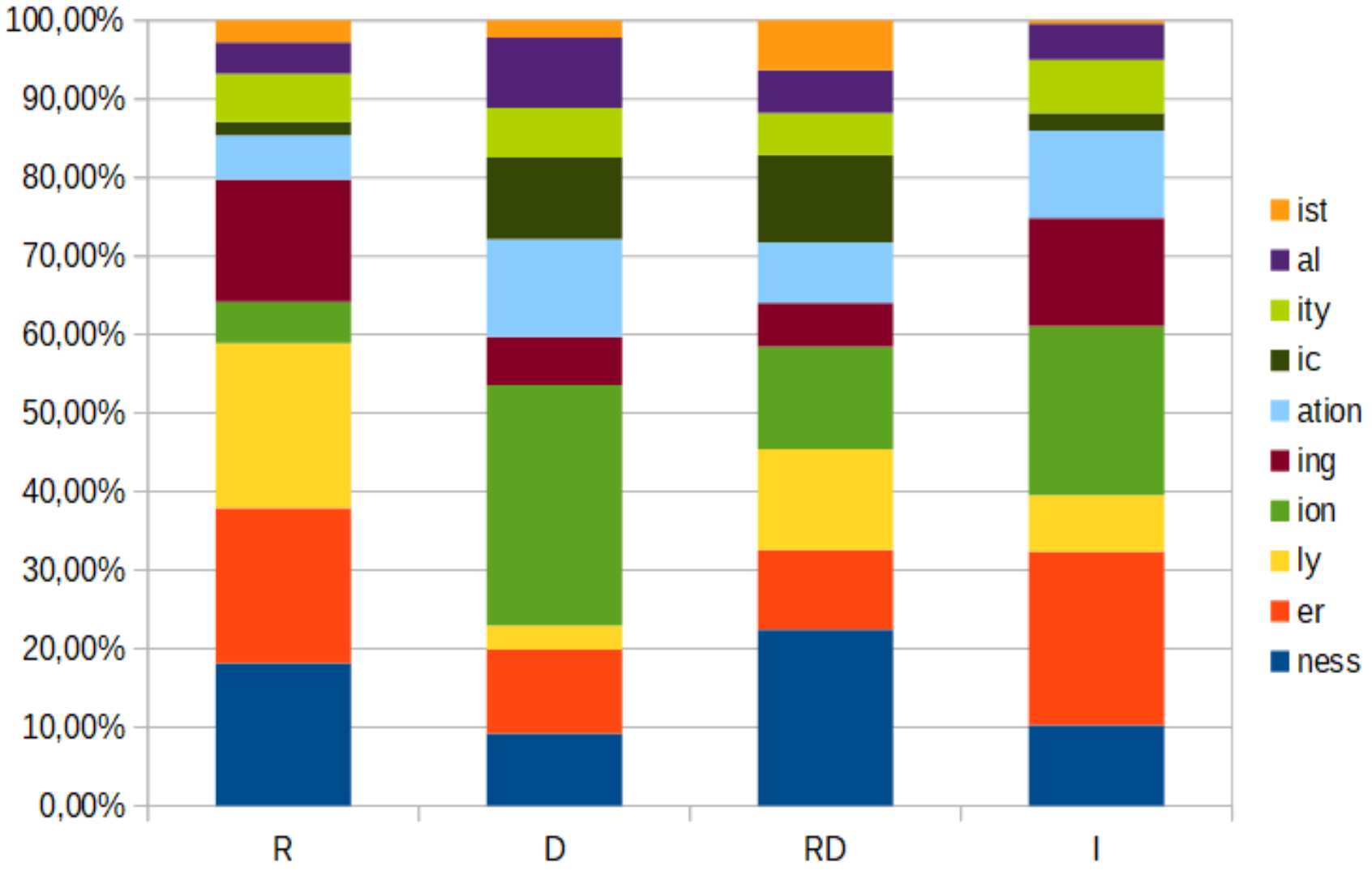
The 10 most frequent suffixes

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

Affixes and derivation types



Derivation types and affixes



Affixes productivity

- We compared our results to those reported by Hay and Baayen (2002), based on a set of words extracted from the CELEX Lexical Database, and noted the correlation of the results.
- Similarity in terms of affixes frequency: -er, -ly, -y, -ness, -al, -ic, -ity, -able
  - Hay and Baayen: + -like, -less
- Hapax legomena (CELEX) – derivation types D and I (PWN): -er, -y, -ly, -ness, -or, -able, -an

Conclusions

- Tested hypothesis: affixes that are involved in deriving words that develop meanings independently from their base word are morphologically productive ones
- The results are biased by several factors:
  - Polysemy – homography
  - the degree of coverage and of correctness of the derivational links in PWN varies from one affix to the other
  - Their impact on the research was not evaluated

Further work

- check if PWN granularity is reflected in the way derivation is marked: for each derived literal check the number of derivational links each of its senses establishes with the base word
- 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)
- Affixes capacity of allowing sense evolution (calculated as the percent of senses specific to the derived word)
- The semantic types of the base words to which one affix can attach

Acknowlegement

Special thanks to **Harald R. Baayen** for the fruitful discussions during this experiment.