# Competition in evaluation: a multifactorial study of Italian intensifying prefixes

Ivan Lacić

Alma Mater Studiorum – Università di Bologna

Morphology Workgroup, Laboratoire de Linguistique Formelle
Paris, November 8, 2024


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Talk Structure

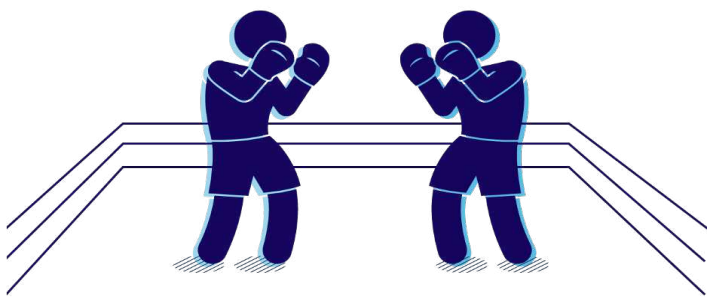
**Theoretical contextualization**

**Case-study presentation**

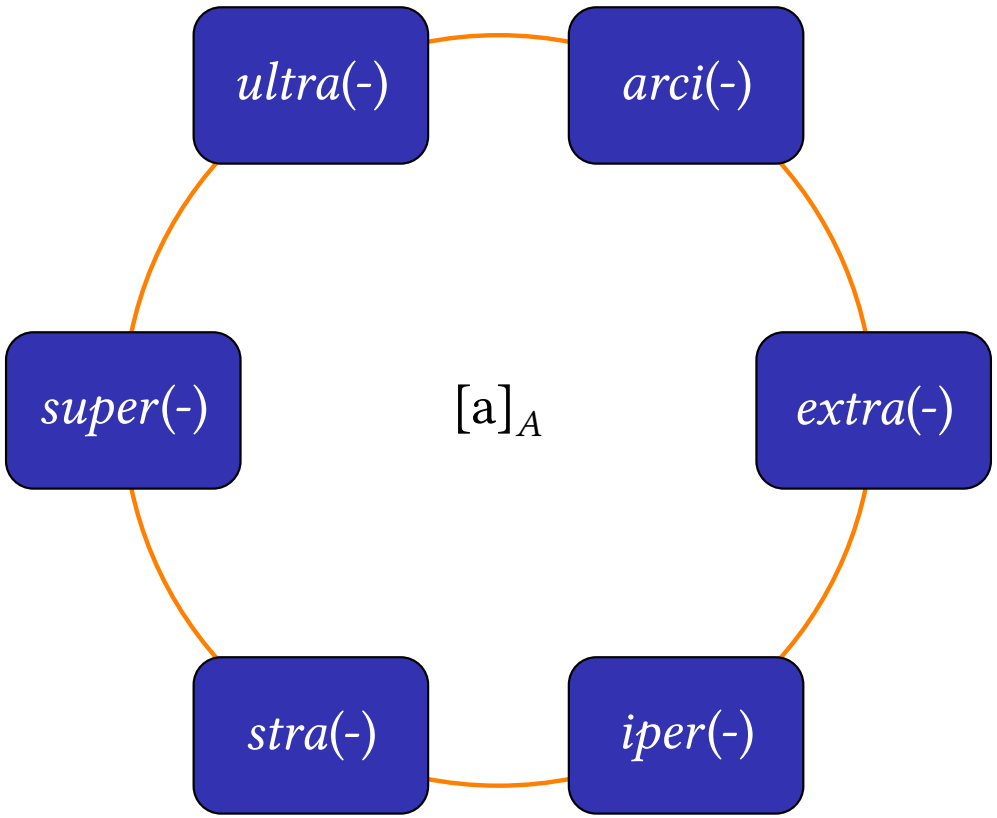
# Theoretical contextualization

# Affix rivalry

- the relationship between two or more affixes that, in at least some of their uses (gradient phenomenon), can form words of identical or similar semantic types (Huyghe and Varvara 2023; Guzmán Naranjo and Bonami 2023)
- under the traditional view (cf. Bréal's law of differentiation), no two linguistic forms with the same function can persist in language → resolution of rivalry
- however, as Nagano, Bagasheva, and Renner 2024 point out, in W-F the competition often simply continues since "(i) coexistence rather than disappearance is commonly observed, and (ii) the form of specialization tends to deviate from the elsewhere distribution [cf. Aronoff 2023]."
- "Quantitative approaches are particularly suitable for the description of affix rivalry, given both its inherent gradience and the multiplicity of factors that can be involved in its resolution." (Huyghe and Varvara 2023)



# Case-study presentation




$$[[\textsc{intens}](\text{-})[\text{a}]_{A,i}]]_{A,j} \leftrightarrow [\text{intens SEM}_i]_j$$

$$[[\textsc{intens}][\text{a}]_{A,i}]_{AP,j} \leftrightarrow [\text{intens SEM}_i]_j$$

Most processes are associated with multiple contents and most derivational relations can be realized by multiple processes → intensification seems a field in which all six prefixes compete

| | locative spatial | locative non spatial | anomalous, strange | superiority in hierarchy | superiority (more than normal) | excess (more than normal) | intensification |
|---|---|---|---|---|---|---|---|
| **extra(-)** | *extraurbano*, "suburban"; *extrauterino*, "ectopic" | *extragiudiziale*, "extrajudicial"; *extra-bilancio*, "non-budgetary" | *extracorrente* [di apertura/chiusura], "inrush current"; *extrasistole*, "extrasystole" | | *extra vergine*, "extra vergin" | *extraprofitto*, "extra profit" | *extrafino*, "extra-fine"; *extrasottile*, "extra-thin" |
| **ultra(-)** | *ultraterreno*, *ultramondano*, "otherworldly" | *ultravioletto*, "ultraviolet"; *ultracentenario*, "ultra-centenarian" | | | | *ultranazionalista*, "extreme nationalist" | *ultramoderno*, "ultra-modern"; *ultrapiatto*, "super-slim" |
| **arci(-)** | | | | *arciprete*, "archpriest" | | | *arcinoto*, "very well known" |
| **iper(-)** | | | | | *ipercritico*, "hypercritical"; *ipermercato*, "superstore" | *iperattivo*, "hyperactive"; *ipertiroidismo*, "hyperthyroidism" | *iperveloce*, "hyperfast"; *ipermoderno*, "ultra modern" |
| **super(-)** | *superstruttura*, "superstructure" | *supersonico*, "superstonic" | | *superspecie*, "superspecies" | *superumano*, "superhuman" | *superalimentazione*, "overfeeding" | *superbianco*, "super-white" |
| **stra(-)** | *straripare*, "overflow" | *stragiudiziale*, "extrajudicial" | | | | *strafare*, "overdo; go overboard" | *stragrande* [maggiornaza], "vast majority" |

According to dictionaries *Treccani*, *Il Nuovo De Mauro*, lo *Zingarelli 2024*, il *Sabatini Coletti*

# Examples itWaC

1. *I contratti di sponsorizzazione con marchi **arcinoti** come Pepsi, Polaroid.*
   'Sponsorship contracts with very well-known brands such as Pepsi, Polaroid.'
2. *Si rilassava accendendosi una sigaretta **extra forte**.*
   'He would relax by lighting an extra strong cigarette.'
3. *E le banche, infine, rimangono **iper-selettive** quando devono concedere credito.*
   'And banks, finally, remain hyper-selective when they have to grant credit.'
4. *Mentre voi vi facevate i fatti miei augurandomi del male io mi sono sentita **stra-felice**.*
   'While you were minding my business and wishing me harm, I was overjoyed.'
5. *Abbiamo montato un pneumatico **super ribassato** su un diametro cerchio da 17 pollici.*
   'We fitted a super low tyre on a 17-inch rim diameter.'
6. *Bill e Boris sono sempre più amici, e al summit finlandese hanno dato vita all'ennesima conferenza stampa dai toni **ultra-amichevoli**.*
   'Bill and Boris are getting closer and closer, and at the Finnish summit they held yet another ultra-friendly press conference.'

# Data extraction I

- corpus itWaC (Baroni et al. 2009) containing $\sim$ 1.5 billion words
- for each the six prefixes in exam, itWaC was queried for <PREF> adjectival ngrams as well for adjectival formations that start with <PREF–> and <PREF>, in order to account for constructions in three orthographic variants, i.e. univerbation, hyphenated use, and juxtaposition
- after manual verification, 139,213 adjectival derivative tokens distributed in 4584 derivative types and combining with 2683 adjectival bases were individuated

# Data extraction II

| **Prefix** | Tokens | Types | Hapax legomena |
|---|---|---|---|
| *arci* | 1,318 | 117 | 81 |
| *extra* | 75,109 | 722 | 235 |
| *iper* | 9,695 | 988 | 430 |
| *stra* | 20,924 | 342 | 163 |
| *super* | 12,888 | 1,327 | 528 |
| *ultra* | 19,279 | 1,103 | 492 |

Table 1: Statistical Overview of Prefix Usage

# Annotation for semantic function I

- recall that prefixes are polysemous/polyfunctional
- since our focus is solely on intensified derivatives, it is essential to determine which derivatives in the dataset truly express intensification
- dataset was divided into two parts and the annotation was carried out by a total of five annotators, three for each part
- due to the sample size, we annotated types, not tokens

| **Prefix** | RA (%) | Fleiss' $\kappa$ | AC1 | 95% CI |
|---|---|---|---|---|
| *arci* | 98.44 | -0.005 | 0.990 | 0.97-1.00 |
| *extra* | 81.99 | 0.629 | 0.873 | 0.85-0.90 |
| *iper* | 80.66 | 0.323 | 0.860 | 0.84-0.88 |
| *stra* | 97.59 | 0.422 | 0.984 | 0.97-0.99 |
| *super* | 86.63 | 0.339 | 0.901 | 0.88-0.92 |
| *ultra* | 87.39 | 0.603 | 0.908 | 0.89-0.93 |

Table 2: Inter-annotator agreement 1/2

| **Prefix** | RA (%) | Fleiss' $\kappa$ | AC1 | 95% CI |
|---|---|---|---|---|
| *arci* | 98.11 | 0.493 | 0.987 | 0.96-1.00 |
| *extra* | 80.74 | 0.609 | 0.871 | 0.85-0.89 |
| *iper* | 82.93 | 0.311 | 0.884 | 0.86-0.90 |
| *stra* | 94.35 | 0.544 | 0.962 | 0.94-0.99 |
| *super* | 86.42 | 0.345 | 0.908 | 0.89-0.92 |
| *ultra* | 82.31 | 0.518 | 0.881 | 0.86-0.90 |

Table 3: Inter-annotator agreement 2/2

# Annotation for semantic function II

| **Prefix** | Total deriv. | Intensified | Non-intensified | Terminology | Unclear (%) |
|---|---|---|---|---|---|
| *arci* | 117 | 116 | 1 | 0 | 0 (0%) |
| *extra* | 722 | 122 | 592 | 1 | 7 (0.97%) |
| *iper* | 988 | 904 | 37 | 31 | 16 (1.62%) |
| *stra* | 342 | 336 | 6 | 0 | 0 (0%) |
| *super* | 1,327 | 1,244 | 51 | 10 | 22 (1.66%) |
| *ultra* | 1,103 | 964 | 122 | 3 | 14 (1.27%) |
| SUM | 4,599 | 3,686 | 799 | 40 | 59 (1.28%) |

Table 4: Overview of the annotation results

# Annotation for semantic function III

| **Prefix** | Tokens | Types | Hapax legomena |
|---|---|---|---|
| *arci* | 1,297 | 116 | 80 |
| *extra* | 837 | 122 | 43 |
| *iper* | 7,930 | 904 | 398 |
| *stra* | 18,581 | 336 | 158 |
| *super* | 11,167 | 1,244 | 503 |
| *ultra* | 8,257 | 964 | 460 |

Table 5: Statistical Overview of Prefix Usage – Intensifiers Only

# Annotation for semantic function IV

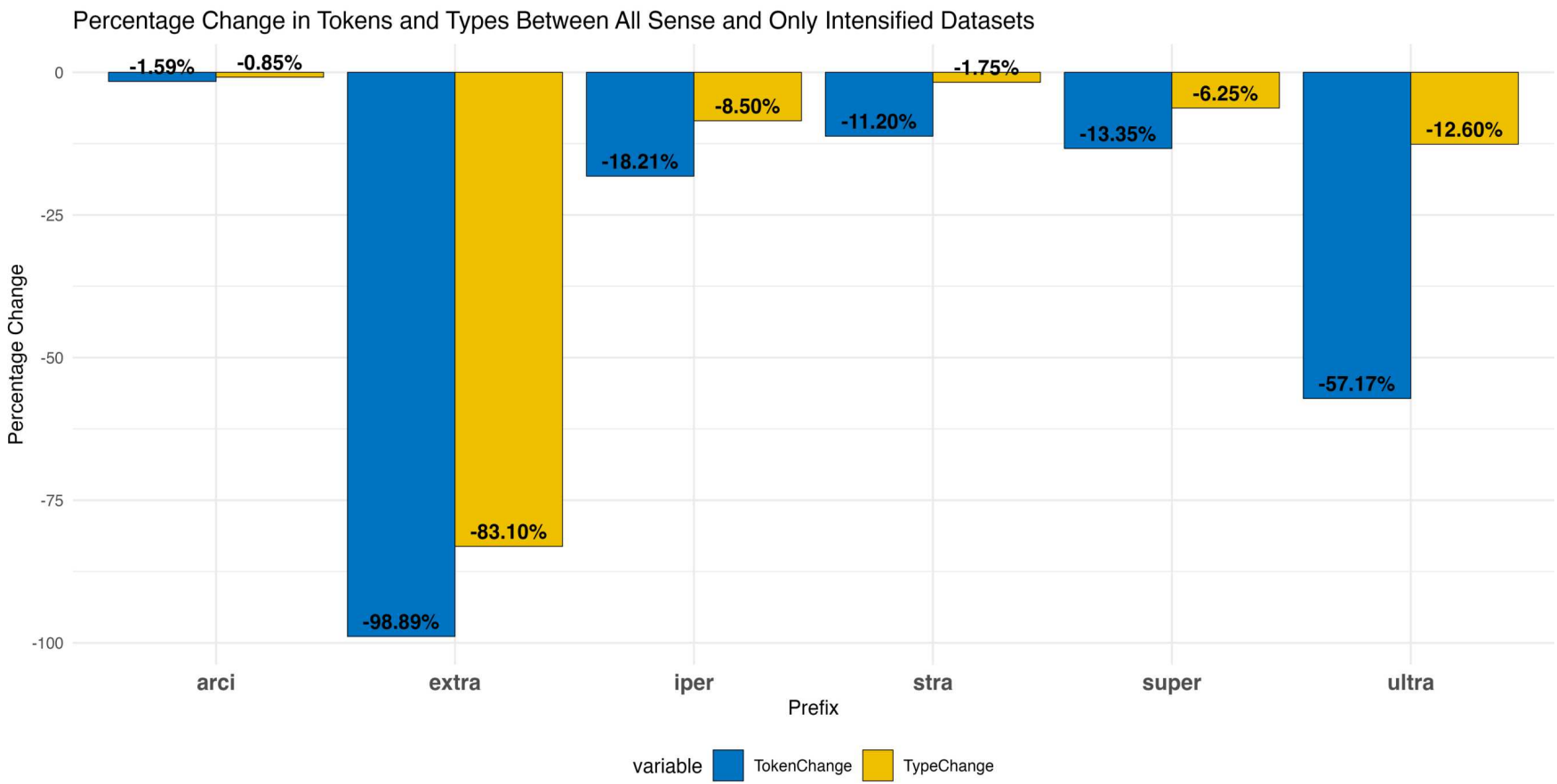
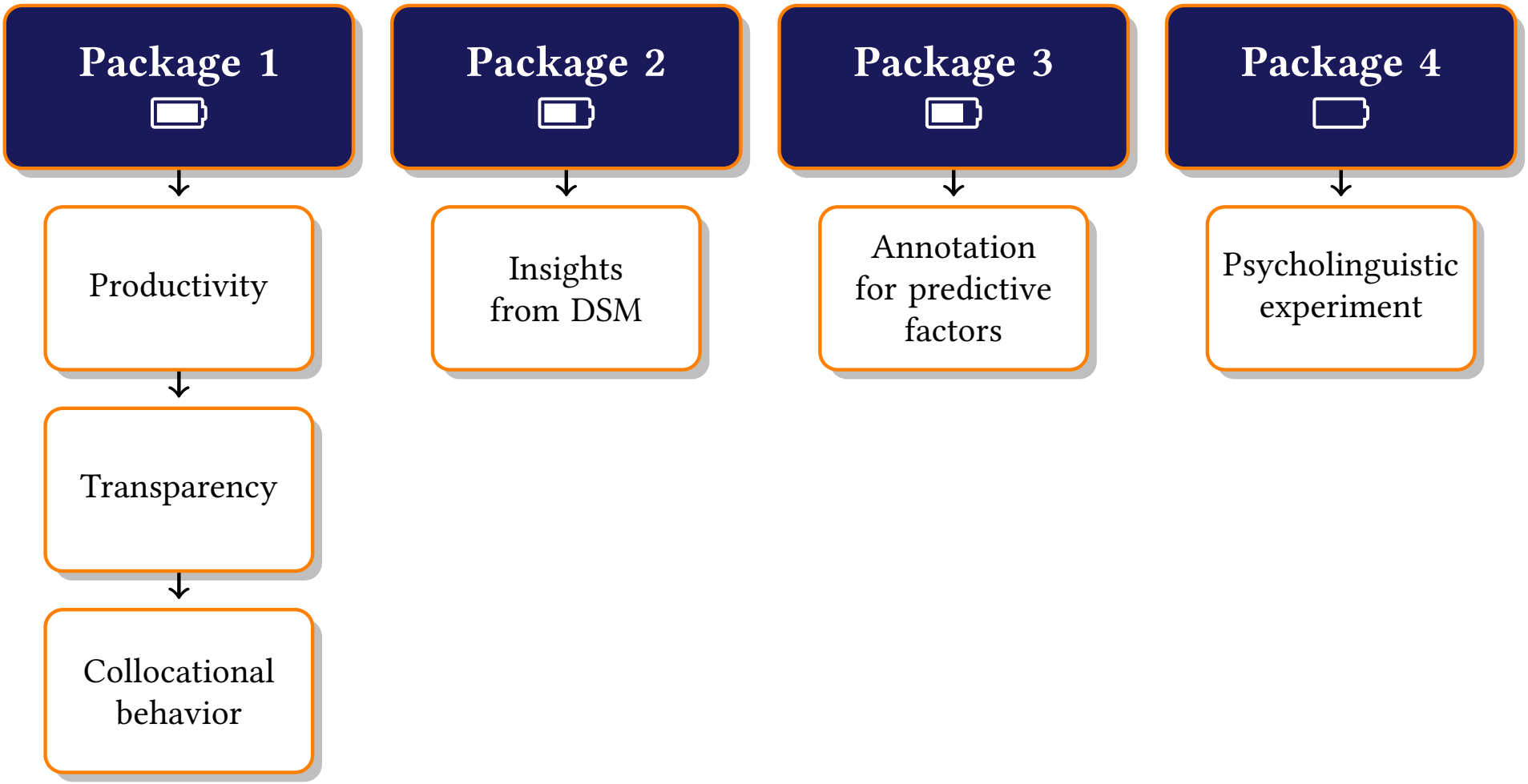
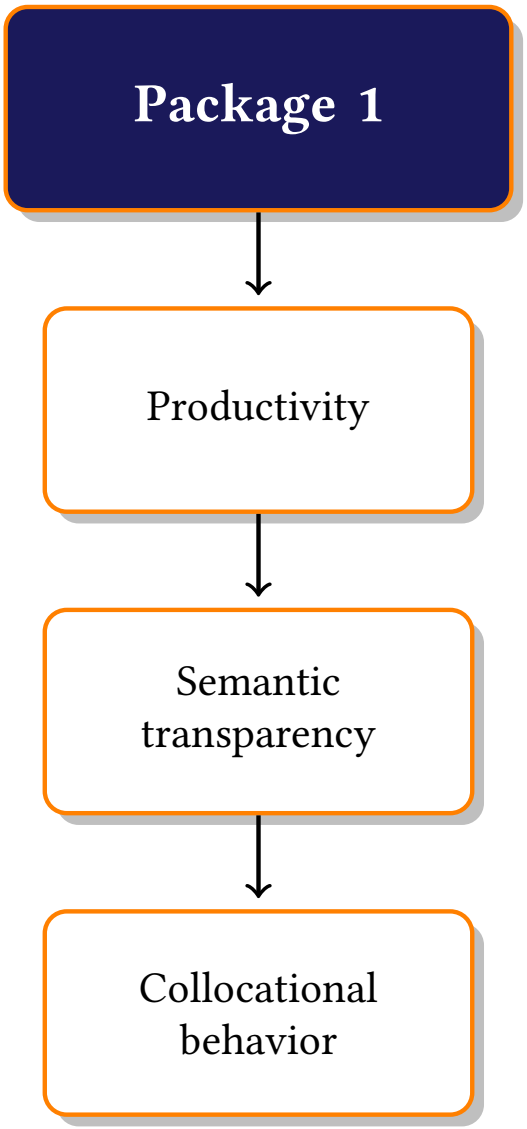



Figure 1: The percentage change in type and token counts for derivatives of all senses and only intensification ones.

# Research Packages and Steps




Package 1
Productivity
Semantic transparency
Collocational behavior

# RSQ1: How productive are the prefixes?

- morphological productivity → phenomenon that a morphological pattern (a systematic form-meaning correspondence) observed a set of complex words can be extended to new cases (Booij and Wouden 2017)
- type count (*rentabilité* / realized productivity) is one of the most straightforward measures of morphological productivity – the greater the number of types associated with a given morpheme, the more productive the morpheme is considered to be (Bauer 2001)
- due to differences in sample sizes across prefixes, comparing raw type counts is not meaningful

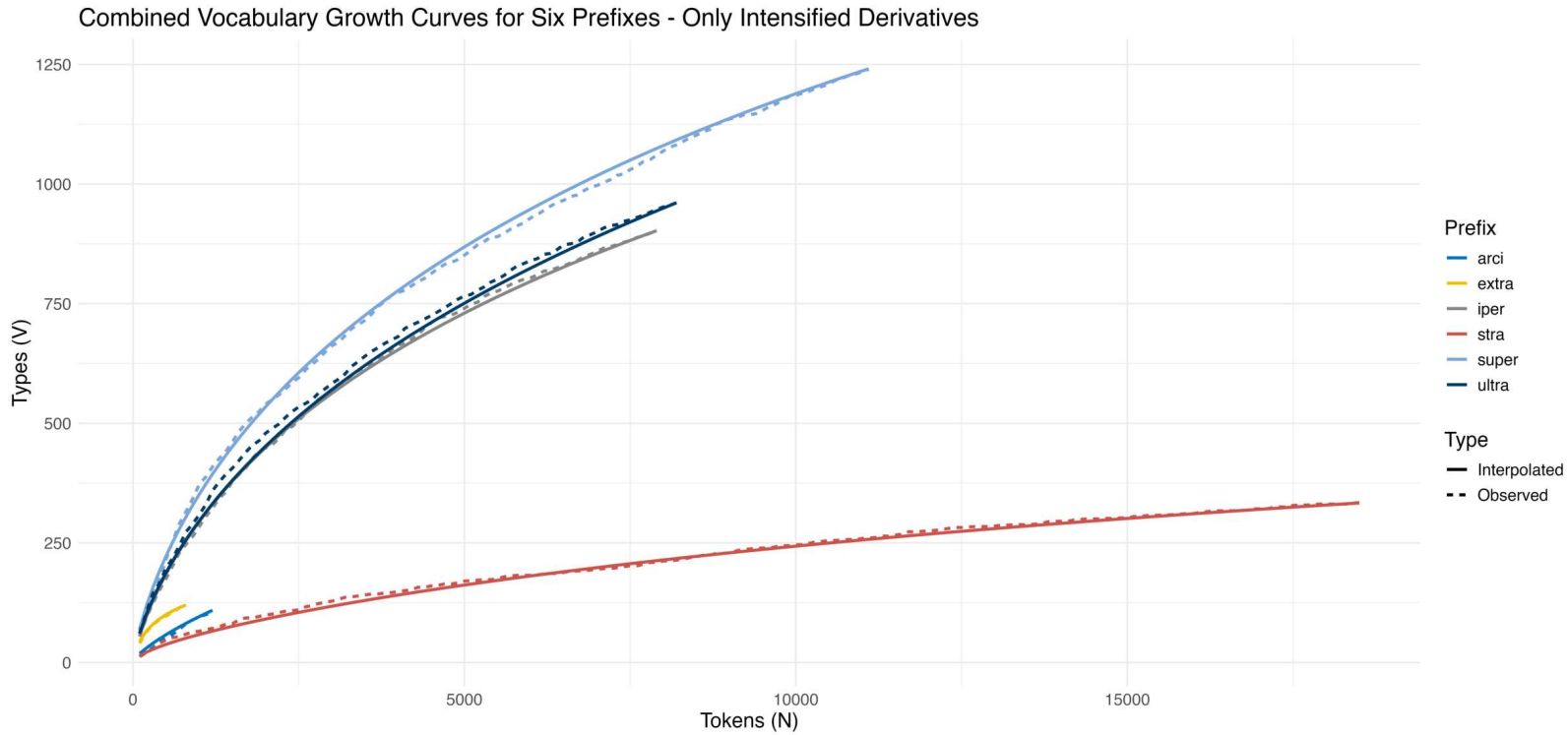
| Prefix | Tokens | Types | Hapax legomena |
|---|---|---|---|
| *arci* | 1,297 | 116 | 80 |
| *extra* | 837 | 122 | 43 |
| *iper* | 7,930 | 904 | 398 |
| *stra* | 18,581 | 336 | 158 |
| *super* | 11,167 | 1,244 | 503 |
| *ultra* | 8,257 | 964 | 460 |

Table 6: Type count per prefix

# RSQ1: How productive are the prefixes?

Vocabulary growth curves

- display the vocabulary size, i.e. the number of types in relation to the increasing number of tokens generated by the examined process
- modeled using the finite Zipf-Mandelbrot LNRE model[1] from `zipfR` (Evert and Baroni 2022)




[1] The goodness of fit results are satisfactory (*arci*: $\chi^2(13) = 1.142$, p = 0.767; *stra*: $\chi^2(13) = 1.148$, p = 0.887; *iper*: $\chi^2(13) = 11.947$, p = 0.289; *ultra*: $\chi^2(13) = 16.089$, p = 0.138; *extra*: $\chi^2(13) = 1.010$, p = 0.799), with the exception of suboptimal fit for *super* ($\chi^2(13) = 36.094$, p = 5.737e-4).

# RSQ1: How productive are the analyzed prefixes?

Four productivity measures:

1. **Type–Token Ratio** ($TTR$)

$$TTR = \frac{types}{tokens} \tag{1}$$

2. **Potential productivity** $\mathcal{P}$ (Baayen and Lieber 1991a; Baayen 2009)

$$\mathcal{P} = \frac{hapaxes}{tokens} \tag{2}$$

3. **Shannon entropy** ($H$) (cf. Hein and Brunner 2019; Evert and Baroni 2022)

$$H(p) = -\sum_i p_i log_2 (p_i) \tag{3}$$

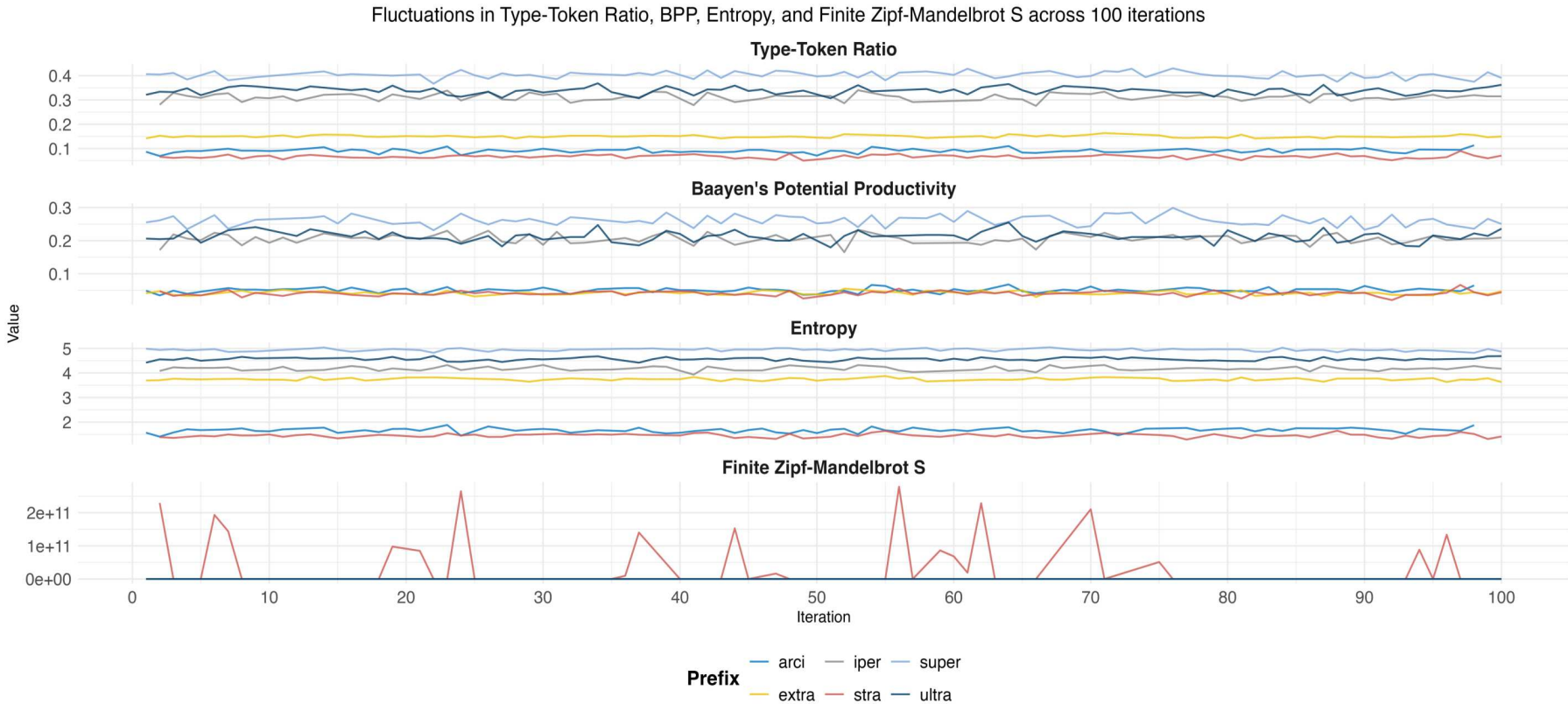
4. **fZM population vocabulary size** ($S$), i.e. the maximum number of types generated by a morphological process; for underlying math see Evert and Baroni 2022

Combining four measures with sample size held constant: 635 token sub-sample of each prefix ($\sim$ 70% of the least freq prefix – *arci* – sample size).
The sampling process is executed 100 times, picking new random samples (*with replacement* function) and calculating the $TTR$, $\mathcal{P}$, $H$, and $S$ scores for each iteration.

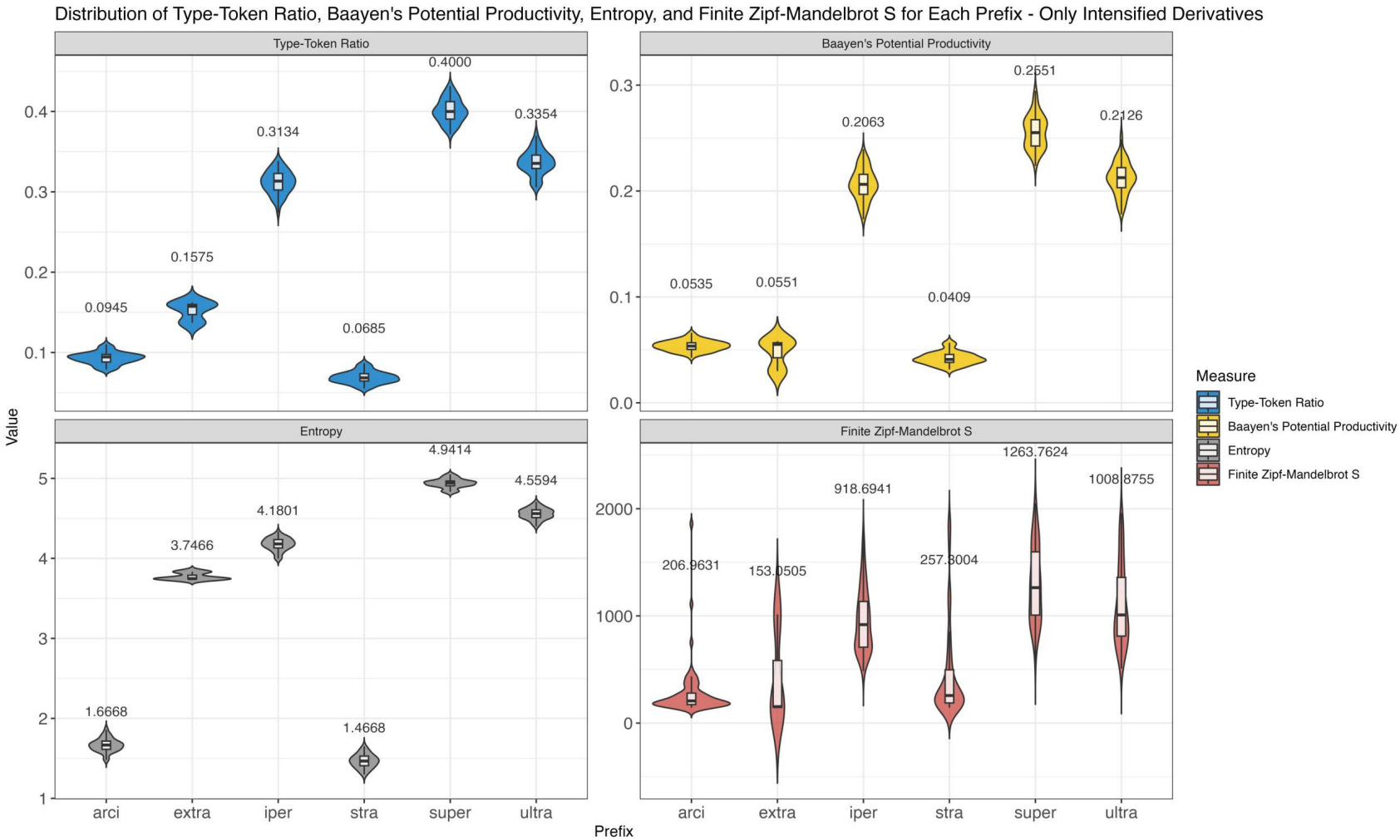
# RSQ1: How productive are the analyzed prefixes?

Fluctuations in productivity values

- min-max normalized Coefficient of Variation across measures indicates $S$ as the most volatile measure (CoV = 1)
- known methodological problem → several attempts were made to resolve it (*m.max* optimization, outlier removal using IQR, etc.) but each of them introduced new complications



# RSQ1: How productive are the analyzed prefixes?




Kruskal-Wallis test: significant differences in $TTR$ ($\chi^2$ = 326, p = 2.65e-68), $\mathcal{P}$ ($\chi^2$ = 301, p = 6.99e-63), $H$ ($\chi^2$ = 331, p = 2.74e-69), and $S$ ($\chi^2$ = 231, p = 7.50e-48) values across prefixes.
Dunn's post-hoc with Bonferroni correction: majority of the pairwise comparisons statistically significant ($p < 0.05$) after correction for multiple testing; certain pairs (*arci-extra*, *extra-iper*, and *iper-ultra*) did not reach statistical significance.

# RSQ1: How productive are the analyzed prefixes?

## Conclusions about productivity

- setting $S$ aside, results of the fixed-size sample analysis are consistent and the rankings according to three productivity measures align perfectly (Pearson's $\rho$ spans from 0.967 for $\mathcal{P}$–H to 0.999 for TTR–H correlation)
- while the adopted measures capture different aspects of productivity, when the fixed-sample size rule is applied, they are highly correlated and can be used interchangeably
- *super* is the most productive intensifying prefix, closely followed by *ultra* and *iper*. On the other hand, *stra* and *arci* are deemed the least productive (highly lexicalized derivatives such as *stragrande* 'vast' and *arcinoto* 'very well-known')
- the findings corroborate the well-established postulate that rival affixes typically exhibit differences in productivity levels (Bybee 1985; Baayen and Lieber 1991a; Plag 1999; Gaeta and Ricca 2015)

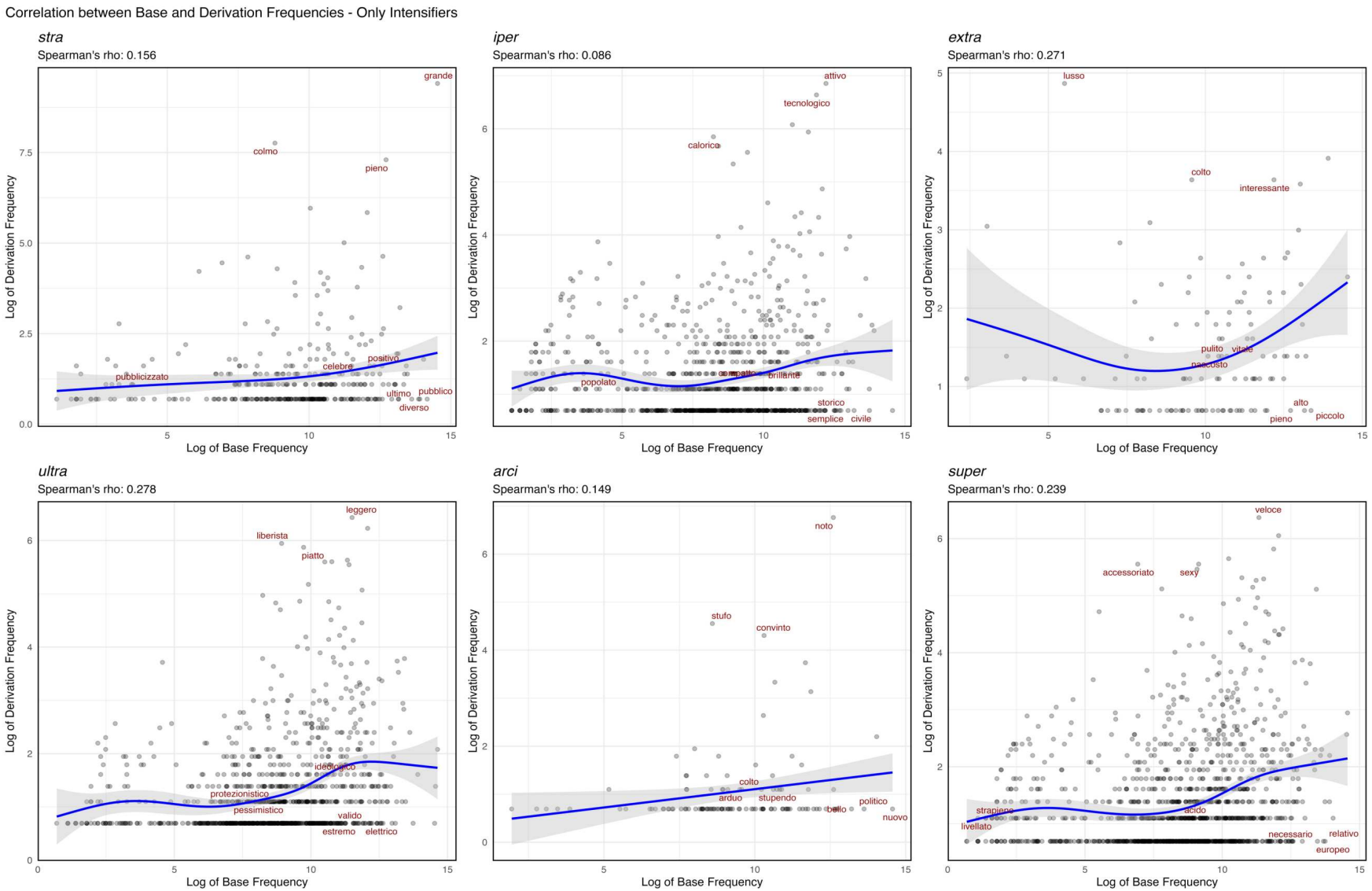
# RSQ2: How semantically transparent are the prefixes? → relative frequency

- semantic transparency → meaning predictability (Plag 2003; Bell and Schäfer 2016) / analyzability (Zwitserlood 1994)
- several works (Bybee 1985; Baayen 1992; Shen and Baayen 2022) make an explicit connection between frequency, semantic transparency, and productivity: high absolute frequency correlates with lower semantic transparency, while lower semantic transparency predicts lower productivity; for a psycholinguistic account see *dual-route theory of processing* (McQueen and Cutler 1998)
- *relative frequency* (Hay and Baayen 2001), i.e. the difference between the frequency of a derivative and the frequency of its base, is understood as having the utmost importance when it comes to semantic transparency and decomposability

# RSQ2: How semantically transparent are the prefixes? → relative frequency

- we explore the correlation of frequency of the adjectival base and the corresponding derivative for each of the six prefixes
- ln of the absolute frequencies is used since speakers tend to process frequency in a logarithmic manner (Hay and Baayen 2001)
- relative frequency effects are visualized with Generalized Additive Modeling (GAM) (Wood 2017)
- based on the residuals from the fitted GAM, each plot includes five adjectival types: i) more frequent as base words (under GAM curve); ii) more frequent as derivatives (above the GAM curve); iii) equally frequent both standalone and in derivations (approximately on the GAM curve, residuals set to $\leq 0.01$)

# RSQ2: How semantically transparent are the prefixes? → relative frequency



# RSQ2: How semantically transparent are the prefixes? → relative frequency

- prefixes display weakly positive or non-monotonic correlations between base and derivation frequencies, with varying degrees of strength
- in accordance to previous studies, e.g. Baayen and Lieber 1991b; Hay and Baayen 2001, the majority of derived forms occurs less frequent than their bases
- *extra* (39.3%) and *super* (35.5%) have the most data points above the GAM curve, suggesting that derivatives they form are less semantically transparent and more likely accessed as independent lexical entries
- conversely, the two least productive prefixes, viz. *arci* (71.3%) and *stra* (69.2%), have the most data points under the GAM curve, presenting the highest decomposability rate

# RSQ2: How semantically transparent are the prefixes? → $\cos(\vec{B}, \vec{D})$

- an alternative approach based on methods from distributional semantics, as presented in, *inter alia*, Marelli and Baroni 2015 and Varvara, Lapesa, and Padó 2021, can be applied
- 12 different DSM configurations built using the `gensim` implementation of `word2vec` algorithm; after evaluating performances on Multilingual SimLex-999 and WS-353 (Vulić et al. 2020), the Skip-gram architecture over a 5-dimensional window to build 500-dimension vectors was chosen[2]
- threshold of at least 10 tokens per derivative type → 383 intensified derivative types formed with six prefixes were individuated
- for every base **B** and derivative **D**, cosine similarity between the base vectors and the derivative vectors was calculated:

$$\cos(\mathbf{B}, \mathbf{D}) = \frac{\mathbf{B} \cdot \mathbf{D}}{|\mathbf{B}||\mathbf{D}|} = \frac{\sum_{i=1}^{n} B_i D_i}{\sqrt{\sum_{i=1}^{n} B_i^2} \sqrt{\sum_{i=1}^{n} D_i^2}} \tag{4}$$

[2] since we used word embeddings, it was not feasible to apply the more reliable (cf. Varvara, Lapesa, and Padó 2021) InvCL measure (Lenci and Benotto 2012), as it requires a model with interpretable dimensions, i.e., a count-based model without dimensionality reduction.

| **arci** | **extra** | **iper** |
|---|---|---|
| Median: 0.3668 | Median: 0.3598 | Median: 0.3473 |
| *noto* (0.4465) | *fondente* (0.5782) | *liberista* (0.6713) |
| *stufo* (0.4454) | *colto* (0.4957) | *calorico* (0.5843) |
| *contento* (0.4007) | *lucido* (0.4379) | *proteico* (0.5178) |
| *famoso* (0.3854) | *fresco* (0.4286) | *accessoriato* (0.5090) |
| *convinto* (0.3482) | *duro* (0.3807) | *cromatico* (0.5042) |
| *sicuro* (0.2485) | *lungo* (0.3389) | *consumistico* (0.4951) |
| *nuovo* (0.1869) | *disponibile* (0.3004) | *nazionalista* (0.4903) |
| *conosciuto* (0.1806) | *interessante* (0.2744) | *lipidico* (0.4898) |
| | *fino* (0.2240) | *colorato* (0.4890) |
| | *vecchio* (0.1787) | *reattivo* (0.4708) |

| **stra** | **super** | **ultra** |
|---|---|---|
| Median: 0.3587 | Median: 0.3356 | Median: 0.3576 |
| *meritato* (0.5217) | *sexy* (0.6249) | *liberista* (0.6857) |
| *colmo* (0.5033) | *attillato* (0.5788) | *nazionalista* (0.5437) |
| *figo* (0.4576) | *grandangolare* (0.5679) | *regolabile* (0.5387) |
| *mitico* (0.4303) | *palestrato* (0.5441) | *portatile* (0.5222) |
| *contento* (0.4213) | *blindato* (0.5162) | *chic* (0.5167) |
| *potente* (0.4204) | *panoramico* (0.5118) | *conservatore* (0.5138) |
| *maturo* (0.4028) | *sfruttato* (0.4971) | *reazionario* (0.5121) |
| *convinto* (0.3923) | *accessoriato* (0.4932) | *morbido* (0.4964) |
| *maledetto* (0.3893) | *affollato* (0.4858) | *moderno* (0.4858) |
| *carico* (0.3863) | *calorico* (0.4817) | *liberale* (0.4810) |

| **arci** | **extra** | **iper** |
|---|---|---|
| Median: 0.3668 | Median: 0.3598 | Median: 0.3473 |
| *well-known* (0.4465) | *dark* (0.5782) | *libertarian* (0.6713) |
| *fed up* (0.4454) | *educated* (0.4957) | *caloric* (0.5843) |
| *satisfied* (0.4007) | *shiny* (0.4379) | *protein* (0.5178) |
| *famous* (0.3854) | *fresh* (0.4286) | *equipped* (0.5090) |
| *convinced* (0.3482) | *hard* (0.3807) | *chromatic* (0.5042) |
| *safe/sure* (0.2485) | *long* (0.3389) | *consumerist*(0.4951) |
| *new* (0.1869) | *available* (0.3004) | *nationalist* (0.4903) |
| *known* (0.1806) | *interesting* (0.2744) | *lipidic* (0.4898) |
| | *fine* (0.2240) | *colored* (0.4890) |
| | *old* (0.1787) | *reactive* (0.4708) |

| **stra** | **super** | **ultra** |
|---|---|---|
| Median: 0.3587 | Median: 0.3356 | Median: 0.3576 |
| *deserved* (0.5217) | *sexy* (0.6249) | *libertarian* (0.6857) |
| *full* (0.5033) | *tight-fitting* (0.5788) | *nationalist* (0.5437) |
| *cool* (0.4576) | *wide-angle* (0.5679) | *adjustable* (0.5387) |
| *mythical* (0.4303) | *muscular* (0.5441) | *portable* (0.5222) |
| *happy* (0.4213) | *armored* (0.5162) | *chic* (0.5167) |
| *powerful* (0.4204) | *panoramic* (0.5118) | *conservative* (0.5138) |
| *mature* (0.4028) | *exploited* (0.4971) | *reactionary* (0.5121) |
| *(convinced)*(0.3923) | *equipped* (0.4932) | *soft* (0.4964) |
| *damned* (0.3893) | *crowded* (0.4858) | *modern* (0.4858) |
| *loaded* (0.3863) | *caloric* (0.4817) | *liberal* (0.4810) |

# RSQ2: How semantically transparent are the prefixes? → $\cos(\vec{B}, \vec{D})$

- *arci* (0.3668) has the highest median cosine similarity value, followed by *extra* (0.3598), *stra* (0.3587), *ultra* (0.3576), *iper* (0.3473), and, finally, *super* (0.3356), with the lowest cosine similarity value
- higher cosine similarity implies that the derivative adjectives are semantically more similar to their base forms. Prefixes with higher cosine similarity (e.g., *arci* and *extra*) indicate that the derivative forms retain more of the base adjective's original meaning
- comparison with residual-based results → prefixes with higher cosine similarity also have a relatively higher percentage of adjectives that occur more frequent as a base, with *extra* as an exception → when the derivative form retains much of the base adjective's meaning (i.e., high cosine similarity), the base form tends to be used more frequently
- possible collinearity between derivative freq and cosine similarity (Zhou et al. 2022) → GAM + linear mixed-effects regression to predict $\cos(\vec{B}, \vec{D})$ from log-transformed derivative freq; the fixed effect of log derivative freq statistically significant ($\beta$ = 0.0235, $t(228)$ = 6.22, p < 0.001), with marginal $R^2$ = 0.0597; **conclusion**: frequency alone explains only a small portion of the variance, aligning with Johnson, Elsner, and Sims 2023, who claim that correlation between low semantic transparency and high freq is only a property of highly polysemous derivatives

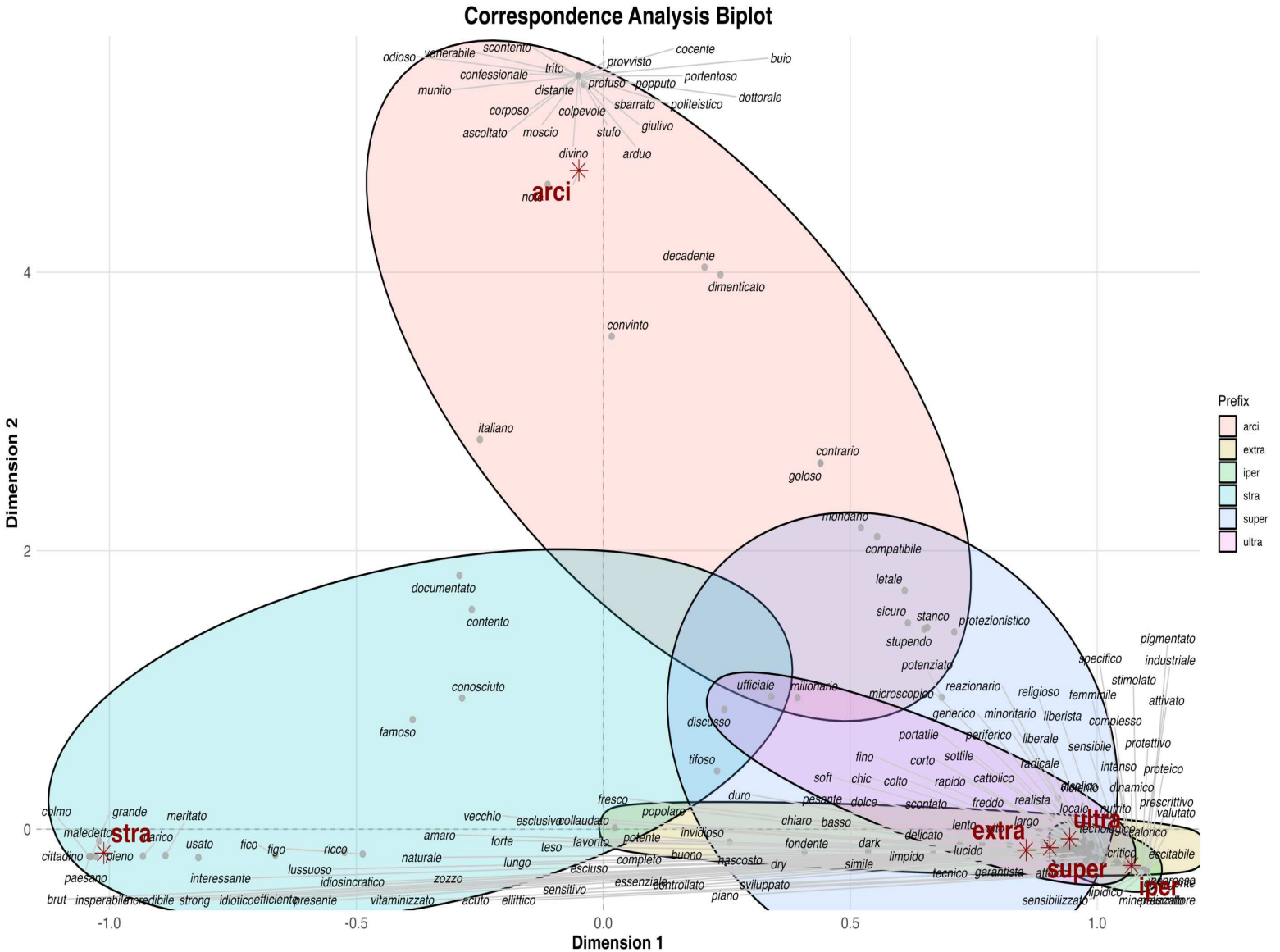
# RSQ3: What is the collocational behavior of the prefixes?

- rival approximative affixes, even when used in analogous contexts, often exhibit distinct distributional tendencies
- to analyze the extent of overlap in collocational preferences between the six prefixes, we employ Multiple Distinctive Collexeme Analysis (MDCA) (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004) to extract the most distinct collexemes of each prefixoids, i.e. the bases that are particularly characteristic of each prefixes
- subsequently, we visualize the most distinct collexemes by means of Correspondence analysis (CA) (Greenacre 2017)

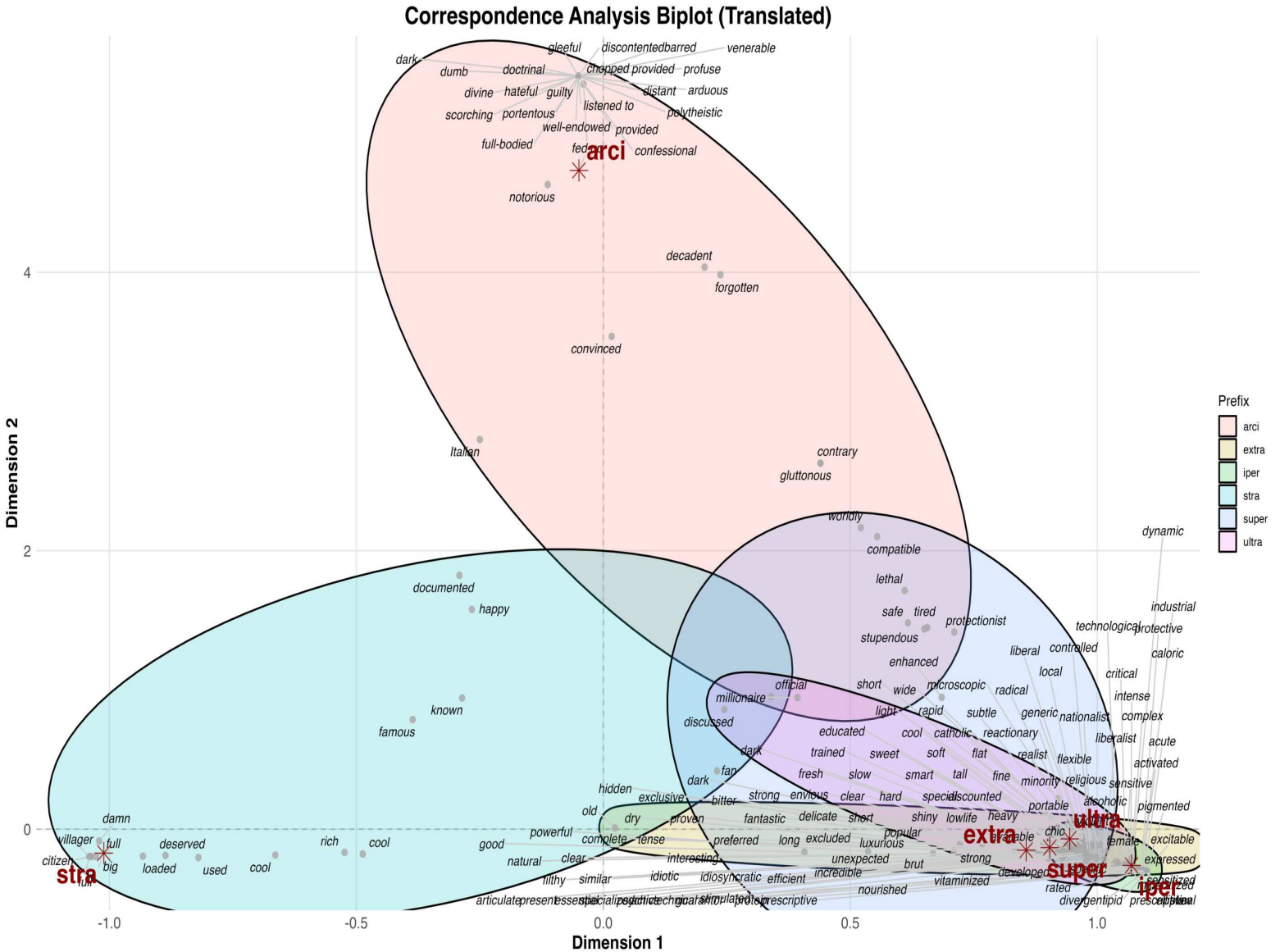
# RSQ3: What is the collocational behavior of the prefixes? → CA

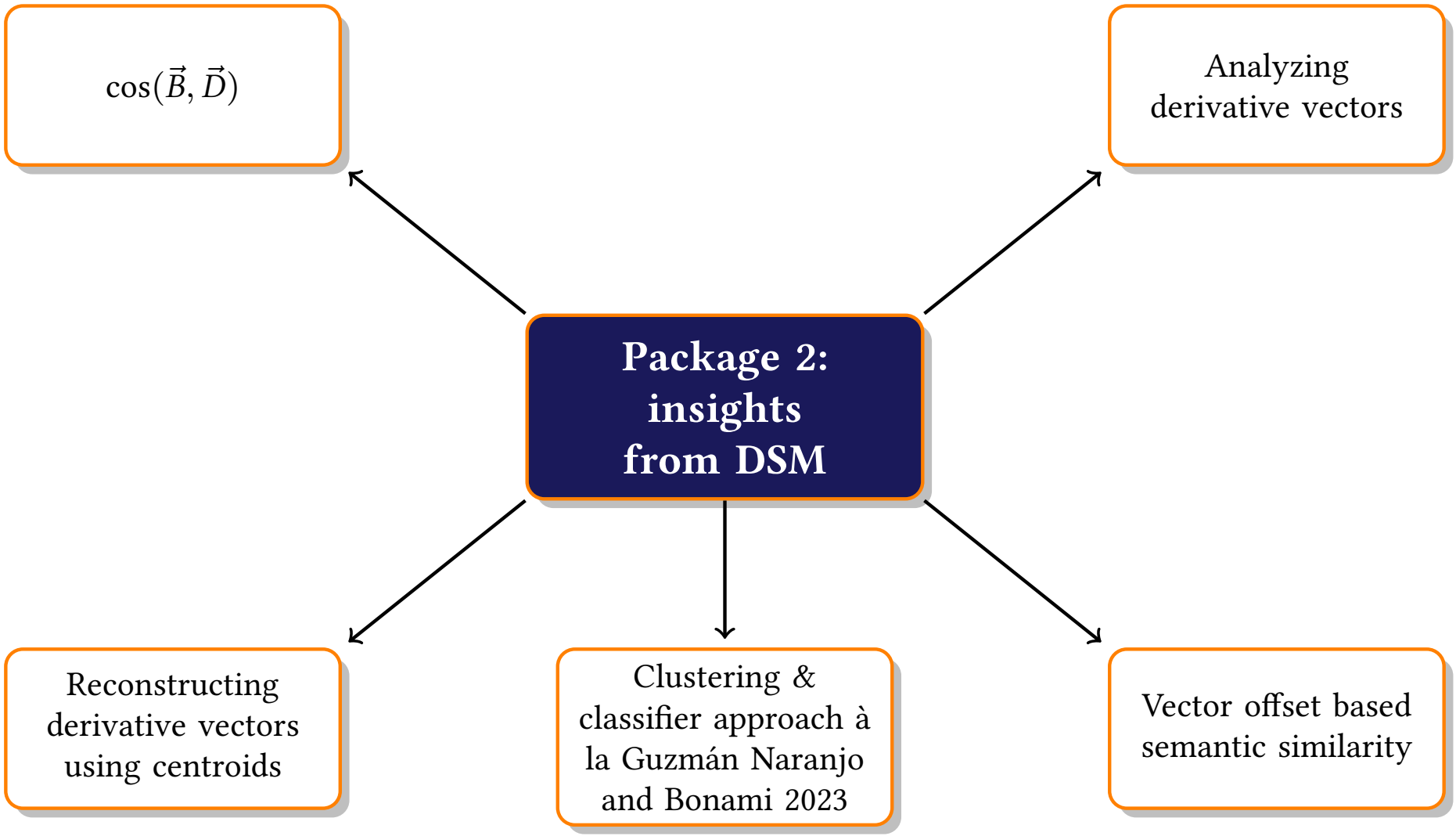
- rows and columns of a contingency table are represented as points in Euclidean space, with their proximity indicating the strength of association
- the $\chi^2$ distance – distance measure akin to Euclidean distance but weighted by the inverse of the average row profile – measures differences between profiles, positioning rows and columns with similar counts closer together.
- conducted using the 50 most distinctive collexemes of each of the six prefixes and the raw frequency of each construction as input
- CA uses the input frequencies to juxtapose (a) line profiles, i.e. distinctive collexemes (adjectives); (b) column profiles, i.e. prefixes; (c) line profiles *and* column profiles, i.e. adjectives *and* prefixes
- the hypothesis of independence regarding the input data can be rejected, with $\chi^2$ = 131,739, p-value = 0; Cramér's $V$ = 0.810 → significant association between the rows and the columns, supporting the notion of a meaningful relationship between the prefixes and adjectives they combine with

# RSQ3: What is the collocational behavior of the prefixes?
# → CA



# RSQ3: What is the collocational behavior of the prefixes? → CA



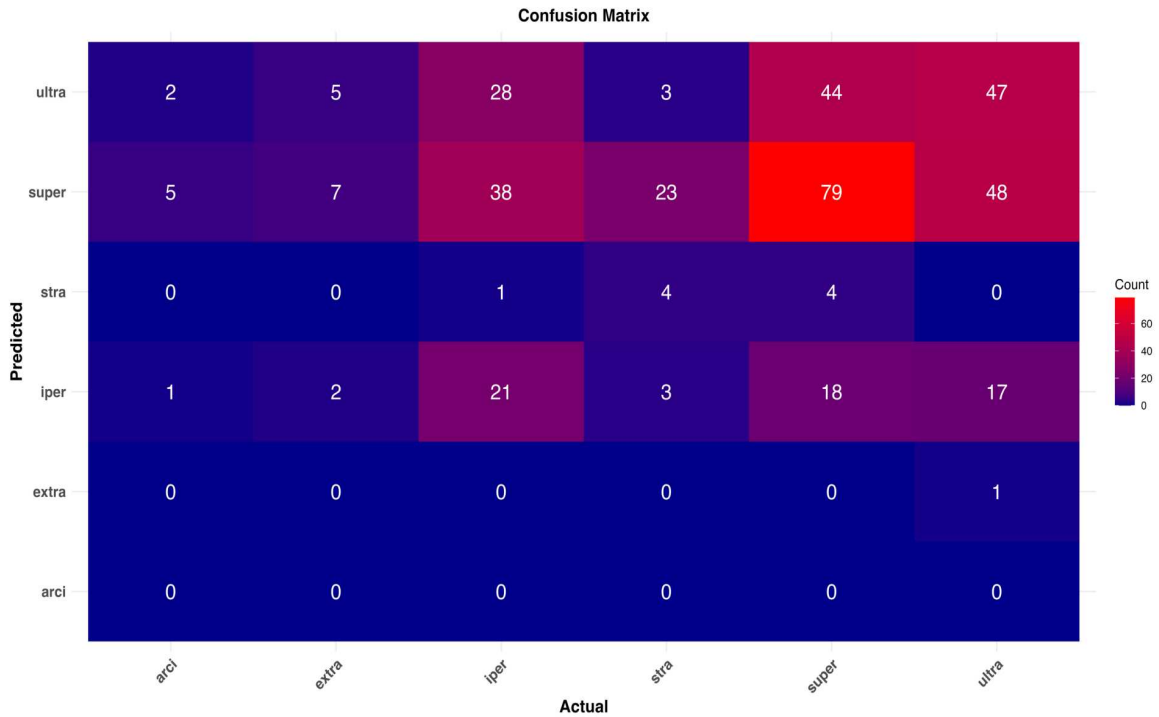

$\cos(\vec{B}, \vec{D})$
Analyzing derivative vectors
Package 2: insights from DSM
Reconstructing derivative vectors using centroids
Clustering & classifier approach à la Guzmán Naranjo and Bonami 2023
Vector offset based semantic similarity

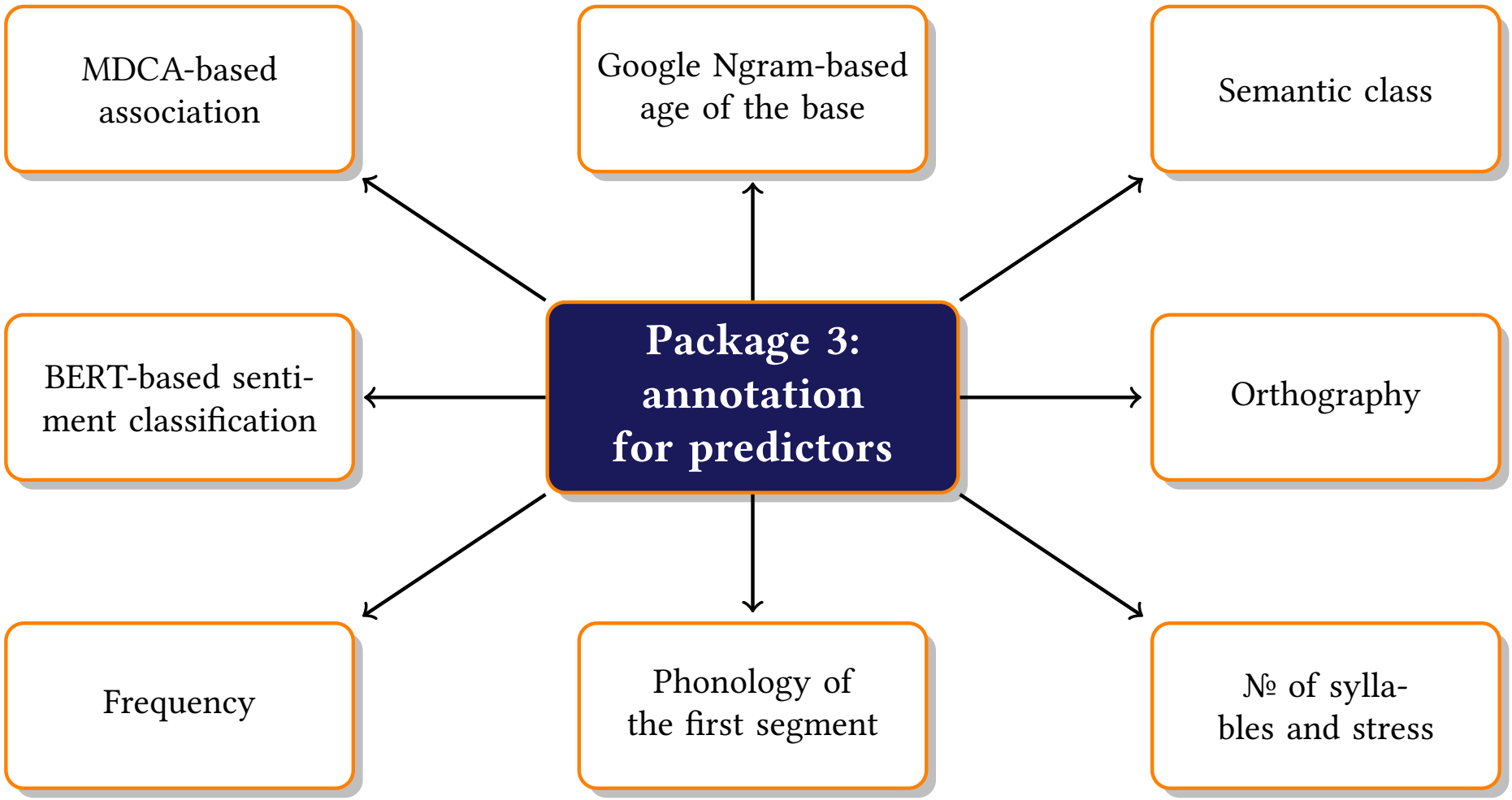
# Package 2: Insights from DSM → vector-offset based classification

- following Guzmán Naranjo and Bonami 2023, we explore whether the semantic information of derivational processes is captured in the distributional vectors of the derivatives
- we apply a machine-learning classification approach in order to determine whether a computational model could reliably classify derivatives based on their prefixes using their respective difference vectors
- if the classifier performs (well) above chance, it is an indication that the vector offsets contain enough semantic information (encode in a satisfying manner the contribution of the prefix) to be able to distinguish one prefix from another
- experiment done using `XGBoost` algorithm (Chen and Guestrin 2016) on dimensionality reduced vectors (PCA, 100 dim, 83.48% variance retained) with 10-fold cross-validation (8 folds for *arci*)
- classifier was trained to predict the prefix that relates the derivative and the base word whose vectors we are comparing based on these difference vectors

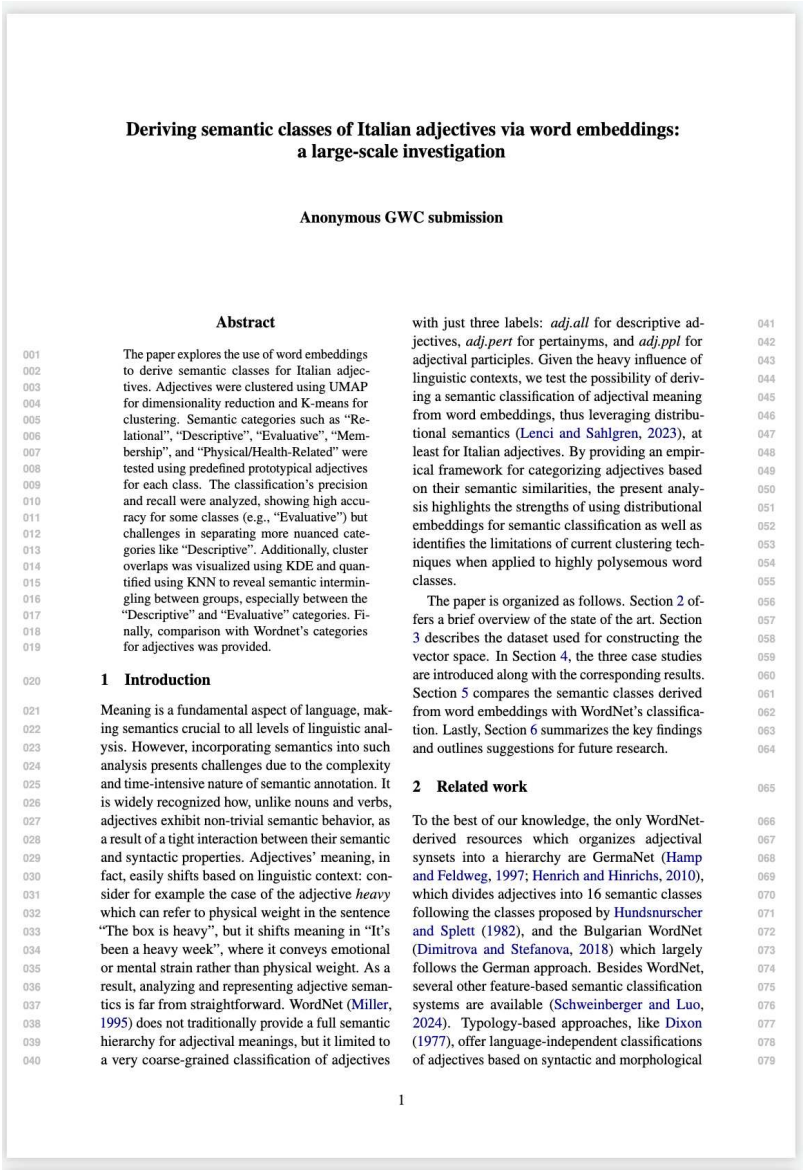
# Package 2: Insights from DSM → vector-offset based classification

- we report the aggregated results of the 10 (8 for *arci*) models on all the left-out data
- classifier achieved an overall accuracy of 37.66% (McNemar's $\chi^2$ of 33.006, p = 9.189e-09), with NIR of 36.16% → **at a chance level** = not capable of distinguishing between different prefixes based on the reduced difference vectors → processes can be considered rivals




MDCA-based association
Google Ngram-based age of the base
Semantic class
BERT-based sentiment classification
Package 3: annotation for predictors
Orthography
Frequency
Phonology of the first segment
№ of syllables and stress

# Package 3: Annotation for predictors → semantic class



**Deriving semantic classes of Italian adjectives via word embeddings: a large-scale investigation**

**Anonymous GWC submission**

**Abstract**

The paper explores the use of word embeddings to derive semantic classes for Italian adjectives. Adjectives were clustered using UMAP for dimensionality reduction and K-means for clustering. Semantic categories such as "Relational", "Descriptive", "Evaluative", "Membership", and "Physical/Health-Related" were tested using predefined prototypical adjectives for each class. The classification's precision and recall were analyzed, showing high accuracy for some classes (e.g., "Evaluative") but challenges in separating more nuanced categories like "Descriptive". Additionally, cluster overlaps was visualized using KDE and quantified using KNN to reveal semantic intermingling between groups, especially between the "Descriptive" and "Evaluative" categories. Finally, comparison with Wordnet's categories for adjectives was provided.

## 1 Introduction

Meaning is a fundamental aspect of language, making semantics crucial to all levels of linguistic analysis. However, incorporating semantics into such analysis presents challenges due to the complexity and time-intensive nature of semantic annotation. It is widely recognized how, unlike nouns and verbs, adjectives exhibit non-trivial semantic behavior, as a result of a tight interaction between their semantic and syntactic properties. Adjectives' meaning, in fact, easily shifts based on linguistic context: consider for example the case of the adjective *heavy* which can refer to physical weight in the sentence "The box is heavy", but it shifts meaning in "It's been a heavy week", where it conveys emotional or mental strain rather than physical weight. As a result, analyzing and representing adjective semantics is far from straightforward. WordNet (Miller, 1995) does not traditionally provide a full semantic hierarchy for adjectival meanings, but it limited to a very coarse-grained classification of adjectives with just three labels: *adj.all* for descriptive adjectives, *adj.pert* for pertainyms, and *adj.ppl* for adjectival participles. Given the heavy influence of linguistic contexts, we test the possibility of deriving a semantic classification of adjectival meaning from word embeddings, thus leveraging distributional semantics (Lenci and Sahlgren, 2023), at least for Italian adjectives. By providing an empirical framework for categorizing adjectives based on their semantic similarities, the present analysis highlights the strengths of using distributional embeddings for semantic classification as well as identifies the limitations of current clustering techniques when applied to highly polysemous word classes.

The paper is organized as follows. Section 2 offers a brief overview of the state of the art. Section 3 describes the dataset used for constructing the vector space. In Section 4, the three case studies are introduced along with the corresponding results. Section 5 compares the semantic classes derived from word embeddings with WordNet's classification. Lastly, Section 6 summarizes the key findings and outlines suggestions for future research.

## 2 Related work

To the best of our knowledge, the only WordNet-derived resources which organizes adjectival synsets into a hierarchy are GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), which divides adjectives into 16 semantic classes following the classes proposed by Hundsnurscher and Splett (1982), and the Bulgarian WordNet (Dimitrova and Stefanova, 2018) which largely follows the German approach. Besides WordNet, several other feature-based semantic classification systems are available (Schweinberger and Luo, 2024). Typology-based approaches, like Dixon (1977), offer language-independent classifications of adjectives based on syntactic and morphological

1

# Some conclusions

UPDATING EXISTING THEORIES

- contrary to Grandi 2017, who claims that *stra* is the "most popular prefix in contemporary Italian, in terms of new formations", we show that *stra* is actually quite unproductive
- Calpestrati 2017 states that (i) "in Italian *extra* mainly has an intensifying function" → we find that *extra*'s contribution to intensification is minimal, as most of its occurrences convey spatial meaning; (ii) "*super* is perceived as the least intensive prefix" → consistent with its high productivity in our study since intensifiers lose their impact (strength) when overused (Tagliamonte 2016)


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Merci!

ivan.lacic2@unibo.it

# References I


Aronoff, Mark (2023). "Three ways of looking at morphological rivalry." In: *Word Structure* 16.1, pp. 49–62.

Baayen, Harald (1992). "Quantitative aspects of morphological productivity." In: *Yearbook of morphology 1991*. Springer, pp. 109–149.

— (2009). "43. Corpus linguistics in morphology: morphological productivity." In: *Corpus linguistics. An international handbook*. de Gruyter, pp. 900–919.

Baayen, Harald and Rochelle Lieber (1991a). "Productivity and English derivation: A corpus-based study." In: *Linguistics* 29, pp. 801–844.

— (1991b). "Productivity and English word-formations: a corpus-based study." In: *Linguistics* 21, pp. 801–843.

Baroni, Marco et al. (2009). "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora." In: *Language resources and evaluation* 43, pp. 209–226.

Bauer, Laurie (2001). *Morphological productivity*. Cambridge University Press.

Bell, Melanie J and Martin Schäfer (2016). "Modelling semantic transparency." In: *Morphology* 26, pp. 157–199.

# References II



Booij, Geert and Ton van der Wouden (2017). *Morphological productivity*. URL: https://taalportaal.org/taalportaal/topic/pid/topic-15033092767688360.


Bybee, Joan L (1985). *Morphology: A study of the relation between meaning and form*.


Calpestrati, Nicolò (2017). "Intensification strategies in German and Italian written language." In: *Exploring Intensification: Synchronic, Diachronic & Cross-Linguistic Perspectives*. John Benjamins, pp. 305–326.


Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system." In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.


Evert, Stefan and Marco Baroni (2022). *Package 'zipfR'*.


Gaeta, Livio and Davide Ricca (2015). "Productivity." In: *Word-Formation: An International Handbook of the Languages of Europe-Vol. 2*. Vol. 40. de Gruyter, pp. 842–858.


Grandi, Nicola (2017). "Intensification processes in Italian." In: *Exploring Intensification: Synchronic, Diachronic & Cross-Linguistic Perspectives*. John Benjamins, pp. 55–77.


Greenacre, Michael (2017). *Correspondence analysis in practice*. Chapman and Hall/CRC.

# References III


Gries, Stefan Th and Anatol Stefanowitsch (2004). "Extending collostructional analysis: A corpus-based perspective onalternations'." In: *International journal of corpus linguistics* 9.1, pp. 97–129.

Guzmán Naranjo, Matías and Olivier Bonami (2023). "A distributional assessment of rivalry in word formation." In: *Word Structure* 16.1, pp. 87–114.

Hay, Jennifer and Harald Baayen (2001). "Parsing and productivity." In: *Yearbook of morphology 2001*. Springer, pp. 203–235.

Hein, Katrin and Annelen Brunner (2019). "Why do some lexemes combine more frequently than others?–An empirical approach to productivity in German compound formation." In: *Mediterranean Morphology Meetings*. Vol. 12, pp. 28–41.

Huyghe, Richard and Rossella Varvara (2023). "Affix rivalry: Theoretical and methodological challenges." In: *Word Structure* 16.1, pp. 1–23.

Johnson, Martha Booker, Micha Elsner, and Andrea D. Sims (2023). "High frequency derived words have low semantic transparency mostly only if they are polysemous." In: *IV International Symposium of Morphology, ATILF Nancy, September 13-15, 2023*.

# References IV


Lenci, Alessandro and Giulia Benotto (2012). "Identifying hypernyms in distributional semantic spaces." In: ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 75–79.

Marelli, Marco and Marco Baroni (2015). "Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics.." In: *Psychological review* 122.3, pp. 485–515.

McQueen, James M. and Anne Cutler (1998). "Morphology in word recognition." In: *he Handbook of morphology*. Blackwell Publishers, pp. 406–427.

Nagano, Akiko, Alexandra Bagasheva, and Vincent Renner (2024). "Towards a competition-based word-formation theory." In: *Competition in Word-Formation*. Benjamins, pp. 1–31.

Plag, Ingo (1999). *Morphological productivity: structural constraints in English derivation*. de Gruyter.

— (2003). *Word-formation in English*. Cambridge University Press.

Shen, Tian and Harald Baayen (2022). "Productivity and semantic transparency: An exploration of word formation in Mandarin Chinese." In: *The Mental Lexicon* 17.3, pp. 458–479.

Stefanowitsch, Anatol and Stefan Th Gries (2003). "Collostructions: Investigating the interaction of words and constructions." In: *International journal of corpus linguistics* 8.2, pp. 209–243.

Tagliamonte, Sali (2016). *Teen talk: The language of adolescents*. Cambridge University Press.

# References V



Varvara, Rossella, Gabriella Lapesa, and Sebastian Padó (2021). "Grounding semantic transparency in context: A distributional semantic study on German event nominalizations." In: *Morphology* 31, pp. 409–446.


Vulić, Ivan et al. (2020). "Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity." In: *Computational Linguistics* 46.4, pp. 847–897.


Wood, Simon N. (2017). *Generalized additive models: an introduction with R*. CRC Press.


Zhou, Kaitlyn et al. (2022). "Problems with cosine as a measure of embedding similarity for high frequency words." In: *arXiv preprint arXiv:2205.05092*.


Zwitserlood, Pienie (1994). "The role of semantic transparency in the processing and representation of Dutch compounds." In: *Language and cognitive processes* 9.3, pp. 341–368.