ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

InTens – a dataset of Italian intensified derivatives. Description and application in a productivity study

Ivan Lacić

Fifth International Workshop on Resources and Tools for Derivational Morphology
Fribourg | September 4, 2025

Roadmap

Motivation

InTens creation
- Data extraction
- Data annotation
- Dataset composition

Putting *InTens* to work: A case study on morphological productivity

Concluding remarks

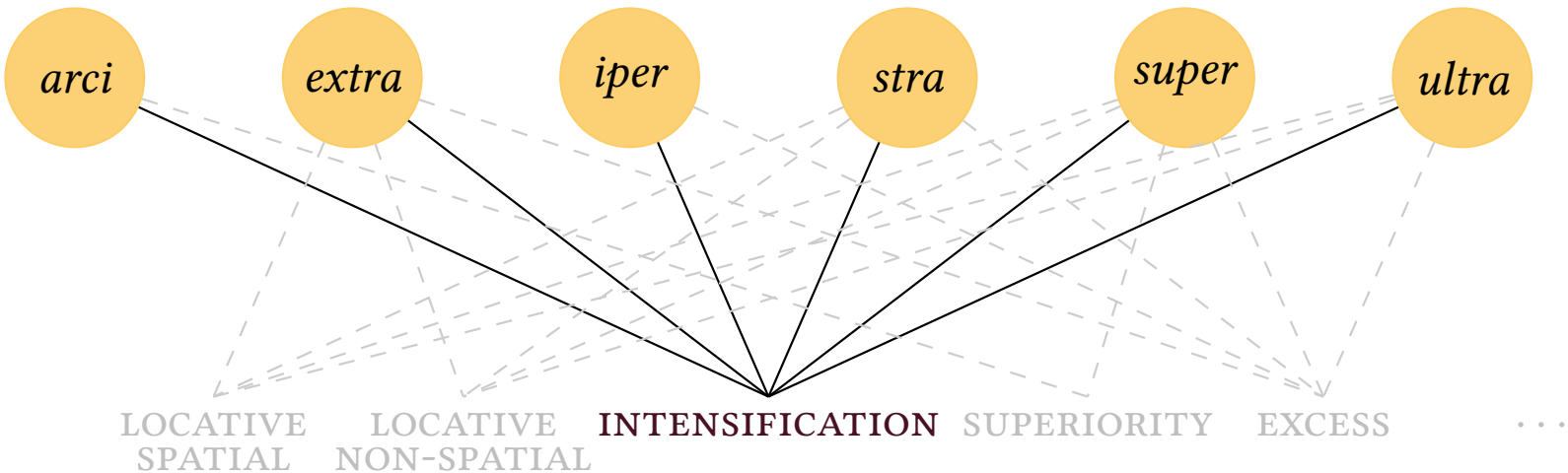
Motivation I

- current resources on Italian word-formation are rather limited (see, e.g., `Morph-It` (Zanchetta & Baroni, 2005) and cf. with `Démonette-2` (Namer et al., 2023))
- we try to make a contribution toward addressing this gap by developing a small-scale resource providing a coherent set of Italian intensified adjectival derivatives named *InTens*
- based on prior literature and numerous exclusion criteria – e.g., removing predominantly quantitative (*maxi* and *mega*) and spatial (*sopra* and *sur*) prefixes – and focusing on prefixes that productively combine with adjectives (cf. Iacobini (2004)), prefixes ***arci***, ***extra***, ***iper***, ***stra***, ***super***, and ***ultra*** were identified[1]
- the utility of *InTens* becomes clear when analyzing the six prefixes in the perspective of affix rivalry → past works on Italian evaluative prefixes have only cataloged them (Montermini, 2008), treated them in isolation (Napoli, 2012), or contrasted them descriptively (Calpestrati, 2017), without addressing their competitive nature

[1] Although we focus on qualitative evaluation, the boundary between qualitative and quantitative evaluation is not clear-cut (Gaeta, 2010; Napoli & Ravetto, 2017). For this reason, we did not exclude prefixes that can also convey quantitative evaluation, provided they predominantly express qualitative evaluation.

Motivation II

- while the six prefixes can be used to derive words of various semantic functions – LOCATIVE SPATIAL (*extraurbano* 'extraurban'), LOCATIVE NON-SPATIAL (*stragiudiziale* 'extrajudicial'), INTENSIFICATION (*ultramoderno* 'ultra-modern'), SUPERIORITY (*arcivescovile* 'archbishop's'), EXCESS (*strapagato* 'overpaid'), etc. – INTENSIFICATION appears the only function in which all six compete[2]



[2] This is not surprising, as predominantly pragmatic communicative objectives that are met through INTENSIFICATION promote the emergence of rivalry as an acceptable pleonastic feature of the system (Dressler et al., 2019; Merlini Barbaresi & Dressler, 2020).

Motivation III

- since rival affixes are typically distinguished by their productivity (Bybee, 1985; Gaeta & Ricca, 2015), and given that productivity should be assessed in relation to specific semantic functions within a morphological pattern (Kastovsky, 1986; Bauer, 2001), it is essential for our study to quantify productivity using a dataset confined to derivatives expressing INTENSIFICATION[3]

[3] Encompassing Amiot's (2004) subclasses of HIGH DEGREE and EXCESS.

Data extraction I

- data were retrieved from the `itWaC` corpus (Baroni et al., 2009), the largest freely downloadable web corpus of Italian[4]

STEP 1

For derivatives formed with one of the six specific strings, constructions in three orthographic variants (xA, x-A, x A) were extracted.

STEP 2

Since the extraction procedure did not distinguish true derivatives from words merely starting with the target strings (e.g., *stra* in *stradale* 'street$_{\text{REL}}$'), adjectives where the string was part of the base were excluded, cross-referencing with *lo Zingarelli* dictionary (Zingarelli et al., 2023).

STEP 3

As many extracted "bases" were in fact fragments of full words (e.g., *tegico* from *strategico* 'strategic'), the forms were cross-referenced with Italian adjectives in Wiktionary (via `kaikki` machine-readable dictionary (Ylonen, 2022)), excluding those not attested.

[4] Spoken corpora such as `KIParla` (Mauri et al., 2019) were unsuitable due to extremely low frequency of relevant derivatives.

Data extraction II

STEP 4

To minimize noise, bases with a frequency < 5 in the corpus were excluded.

STEP 5

Derivatives where prefixation is anterior to denominal suffixation (e.g., *arcidiocesano* 'archdiocesan' < *arcidiocesi* 'archdiocese') were ruled out, retaining only cases where the prefix is the final morphological addition (Fradin et al., 2008; Baayen, 2009).

STEP 6

In cases where multiple derivational paths are possible (e.g., *supertechnologico* 'supertechnological' could be derived either via suffixation from *supertechnologia* 'supertechnology' or via prefixation from *technologico* 'technological'), derivatives that could plausibly admit derivation via prefixation were retained.

Data extraction III

The complete dataset consists of 4,599 derivative types formed with 2,683 adjectival base types, distributed as follows:

	Tokens	Types	Hapaxes
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: Distribution of derivatives across prefixes.

Annotation for semantic function

- recall that the six prefixes are polyfunctional
- after testing multiple (semi)automatic methods, manual semantic annotation of derivatives was selected as the preferred approach
- annotators had three labels at their disposal: (i) INTENSIFICATION, (ii) NON-INTENSIFICATION, and (iii) TERMINOLOGY (scientific terms)
- in ambiguous readings (e.g., *iperattivo* 'hyperactive' can be understood both as conveying values of INTENSIFICATION and EXCESS), annotators were instructed to default to INTENSIFICATION if plausible
- due to the dataset's large size, it was bisected and annotated by the author along with four PhD-level native speakers of Italian
- each half was annotated by a trio: the author, along with annotators 1 and 2 for the first part, and the author, along with annotators 3 and 4 for the second part
- the annotations were performed on a type-based level[5]

[5] To (at least partially) address the assumption of monosemy adopted here, a random sample of 20 derivative types with a frequency ≥ 10 per prefix was individuated and an analysis on a randomly selected sample of 300 tokens was conducted. No instances of genuine polysemous interpretations, encompassing both intensified and non-intensified readings, were identified.

Trustworthiness of annotations

- after a qualitative analysis of the annotations, the computation of inter-annotator agreement in form of raw agreement (RA) (Goodman & Kruskal, 1959) and Fleiss' κ (Fleiss, 1981) was performed
- owing to the highly homogeneous annotations and the so-called κ paradox, a substantial discrepancy between RA and Fleiss' κ values emerged
- to this end, Gwet's AC1 (Gwet, 2008) was introduced[6]
- high RA ($> 80\%$) and AC1 values (≥ 0.86) attest to substantial inter-annotator agreement (cf. Brezina (2018))

	RA	Fleiss' κ	AC1 [95% CI]
arci	98.4	−0.01	0.99 [0.97−1.00]
extra	85.3	0.60	0.86 [0.83−0.89]
iper	87.5	0.32	0.86 [0.83−0.88]
stra	99.4	0.42	0.98 [0.97−1.00]
super	91.3	0.33	0.90 [0.88−0.92]
ultra	89.8	0.60	0.90 [0.88−0.93]

Table 2: Inter-annotator agreement 1/2.

	RA	Fleiss' κ	AC1 [95% CI]
arci	98.1	0.49	0.99 [0.94−1.00]
extra	80.7	0.60	0.86 [0.82−0.89]
iper	83.9	0.32	0.88 [0.86−0.91]
stra	94.9	0.54	0.96 [0.93−0.99]
super	86.4	0.35	0.90 [0.88−0.92]
ultra	82.5	0.52	0.87 [0.85−0.90]

Table 3: Inter-annotator agreement 2/2.

[6] To sidestep methodological debates about metric choice (cf. Silveira and Siqueira (2023) and Vach and Gerke (2023)), we do not propose AC1 as a replacement for Fleiss' κ, but as a complementary metric offering greater stability against category prevalence effects.

Annotation results

- the subsequent phase in the development of the dataset involved the identification of intensified derivatives
- if two of the three annotators assigned the same tag to a derivative, it was classified accordingly
- if each annotator assigned a distinct tag to the same derivative, it was categorized as UNCLEAR

	Types	INTENSIFICATION	NON-INTENSIFICATION	TERMINOLOGY	UNCLEAR
arci	117	116	1	0	0
extra	722	122	592	1	7
iper	988	904	37	31	16
stra	342	336	6	0	0
super	1,327	1,244	51	10	22
ultra	1,103	964	122	3	14
SUM	4,599	3,686	809	45	59

Table 4: Distribution of derivative types across annotation categories.

InTens composition

- given that TERMINOLOGY and UNCLEAR class collectively represented only 2.3% of the dataset, in order to avoid any reliance on subjective authorial judgment concerning their status, they were excluded from further analysis
- finally, *InTens* is composed as follows:

	Intensified tokens	% of total tokens	Intensified types	% of total types
arci	1,297	98.4	116	99.2
extra	837	1.1	122	16.9
iper	7,930	81.8	904	91.5
stra	18,581	88.8	336	98.3
super	11,167	86.7	1,244	93.8
ultra	8,257	42.8	964	87.4

Table 5: Token and type counts of intensified derivatives, along with the percentage of intensified tokens and types within the total sample.

InTens vs. initial dataset

- for *arci*, token and type reductions are marginal (−1.6% tokens, −0.8% types), while for *stra*, *super*, and *iper* they are relatively small, ranging from −11.2% to −18.2% for token counts, and −1.7% to −8.5% for type counts, confirming their core intensifying role
- conversly, *ultra* and especially *extra* show major drops (*ultra*: −57.2% tokens; *extra*: −98.9%), revealing their dominant non-intensifying uses
- *ultra* retains most of its types (−12.6%), indicating that high-frequency (terminological) types were filtered out, while less frequent intensified ones remained
- these patterns underscore the value of semantic annotation in capturing prefix polyfunctionality and rivalry – crucial distinctions would be lost otherwise
- existing descriptions regarding the intensifying prefixes usage are challenged – e.g., contrary to claims that *extra* primarily acts as an intensifier in Italian (Calpestrati, 2017), our data show its intensifying uses are marginal, echoing trends noted also for French (Izert, 2012; Cartier & Huyghe, 2021)

Productivity as a function of semantics: intro

- while quantifying the general productivity of rival affixes is informative, here *InTens* is used to illustrate the added insights obtained by analyzing productivity for a specific semantic function through an annotated dataset
- specifically, **we contrast the prefixes' productivity within the intensifying domain with their productivity in generating derivatives of all semantic values**

Productivity as a function of semantics: method

- we conducted a quantitative study based on type-token ratio (TTR), potential productivity ($\mathscr{P}$) (Baayen, 2009), entropy (H) (Shannon, 1948), and population vocabulary size (S) from the finite Zipf-Mandelbrot model for LNRE (Evert & Baroni, 2007; Baroni & Evert, 2014), with each measure capturing distinct aspects of word formation and usage[7]
- to avoid confounding effects introduced by varying sample sizes, a fixed number of tokens were randomly chosen for each prefix (1000 tokens for the dataset encompassing all derivatives and 635 tokens for *InTens*, ~ 70% of the smallest sample size)
- to ensure the robustness of the results, the calculation procedure was iterated 100 times, drawing new random samples (`with replacement`) in each iteration

[7] TTR measures balance in usage, $\mathscr{P}$ estimates the likelihood of encountering a new type, H can be seen as a measure of unpredictability in the type-frequency distribution, while S estimates the total number of types that would be observed if the entire population were sampled.

Productivity as a function of semantics: results I

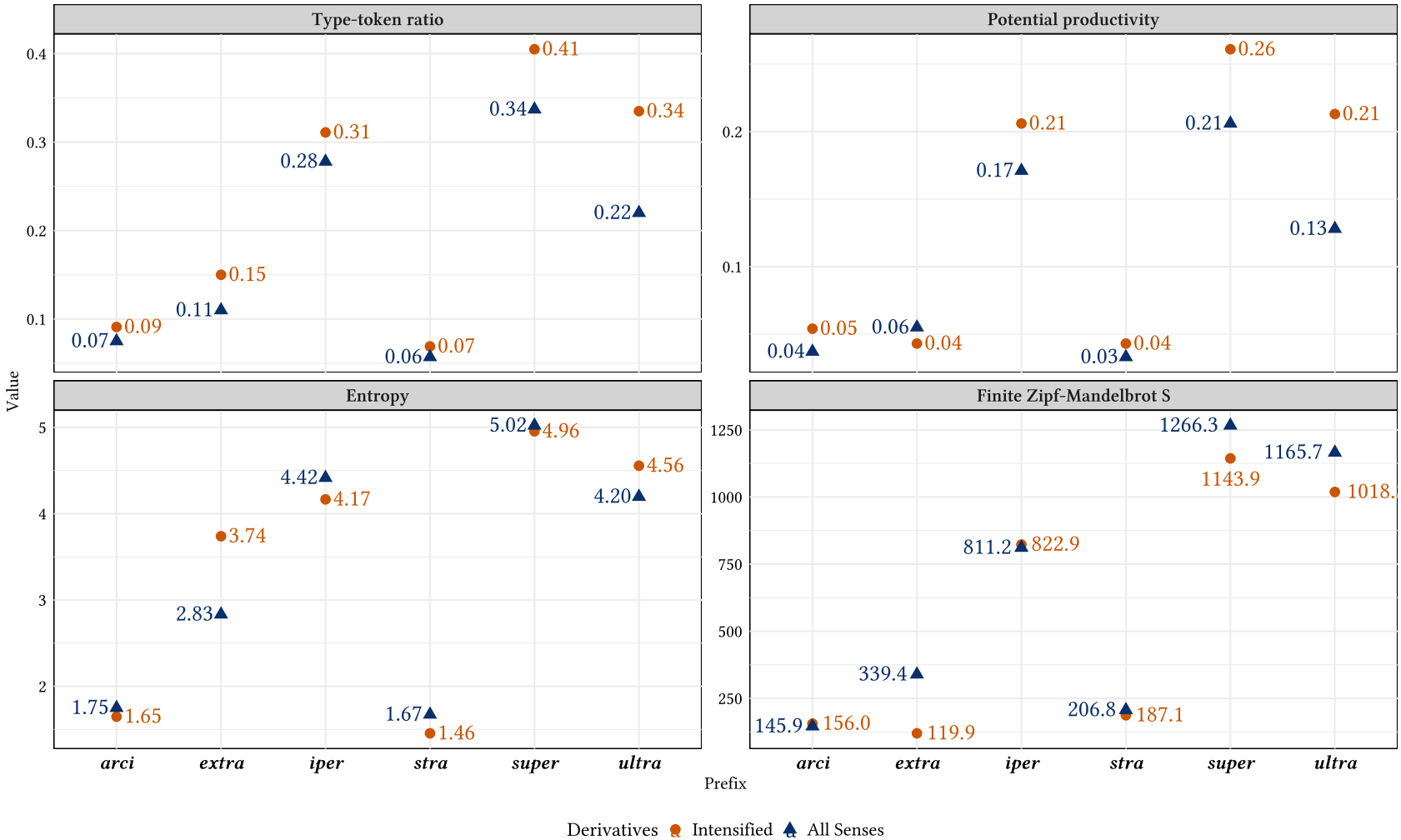


Figure 1: Median productivity values for six prefixes forming derivatives of all semantic functions (represented as triangles) and solely intensified derivatives (represented as dots).

Productivity as a function of semantics: results II

- there is a general trend of increase in TTR and $\mathscr{P}$ when prefixes are used to create intensified derivatives, with the exception of *extra*
- the most significant positive variation in TTR and $\mathscr{P}$ is apparent with *super*
- on the other hand, *stra* and *arci* exhibit minimal changes across all measures, suggesting that their limitation to intensifying contexts does not notably affect their productivity, as this is their principal area of application
- larger discrepancies can be seen in the variations of H and S, with the observed non-uniformity in H variations indicating alterations in the predictability of the derivative distributions associated with each prefix

Productivity as a function of semantics: results III

- in sum, all measures show that *super*, *ultra*, and *iper* are the most productive prefixes, aligning with findings of Cartier and Huyghe (2021) concerning French
- the same findings can be associated with the perception of *super* as the least intense intensifying prefix in Italian (Calpestrati, 2017) → if intensifiers lose their effectiveness from overuse (Mutz, 2015), then *super*'s high productivity and low perceived intensity have a strong correlation
- conversely, *arci* and *stra* exhibit low productivity, likely because a single type dominates much of their samples (65% for *arci* with *arcinoto* 'very well-known' and 58% for *stra* with *stragrande* 'vast') → this supports the common observation that low-productivity processes often include a multitude of high-frequency forms (Plag, 2003)

Productivity as a function of semantics: results IV

- to test whether prefix productivity differs for intensified *versus* all derivatives, and since *InTens* constitutes a subset of the comprehensive "all semantic functions" superset, we opted for a permutation test[8]
- for each prefix, we compared the observed productivity measure M_{obs} from intensified derivatives ($N = 635$) to an empirically constructed null distribution formed by drawing $B = 1000$ bootstrap samples $\{M_i\}_{i=1}^{B}$, each of size $N = 635$ (`with replacement`), from the "all semantic functions" superset
- for each random sample, the empirical two-tailed p-value was calculated as

$$p = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}\left(|M_i - \bar{M}| \geq |M_{\text{obs}} - \bar{M}|\right), \quad \text{where} \quad \bar{M} = \frac{1}{B} \sum_{i=1}^{B} M_i.$$

Here, $\mathbf{1}(\cdot)$ denotes the indicator function, returning 1 if the condition is true and 0 otherwise.

[8] Standard parametric tests (e.g., t-test) were not adequate as they assume independent samples from populations with specific distributional properties, typically normality.

Productivity as a function of semantics: results V

- no significant differences found for *arci*, *iper*, *stra*, *super* ($p_{\text{adj}} = 1$), indicating that restricting the analysis to INTENSIFICATION does not significantly affect their productivity
- strong effects found for *extra* and *ultra* (e.g., *extra* H: $p_{\text{adj}} = 0$, $d = 11.60$; *ultra* TTR: $d = 5.16$; $\mathscr{P}$: $d = 3.64$; H: $d = 6.15$), indicating that the intensified uses of *extra* and *ultra* are significantly more productive than would be expected from their overall derivative pools
- overall, results highlight significant, prefix-specific and measure-specific effects

Summing up

- in this study, we outlined a methodology for building a small-scale dataset of intensified Italian derivatives
- the work highlighted the importance of semantic annotation for polyfunctional prefixes, particularly in the context of affix rivalry, and the illustrative case study demonstrated how the productivity of certain prefixes changed when the domain of interest was narrowed to the semantic function of INTENSIFICATION

Implications and future steps

- *super*, *ultra*, and *iper* exhibit comparably elevated productivity; in contrast, the lower productivity of *arci* and *stra*, along with their ability to generate highly lexicalized derivatives, points to emerging "niche productivity" (Lindsay & Aronoff, 2013)
- a comprehensive understanding of productivity–rivalry dynamics requires large-scale diachronic investigations (cf. Fernández-Domínguez (2017))
- productivity variations may also arise from sociolinguistic, pragmatic, or fashion-driven trends, while new rival processes may emerge due to language change (Merlini Barbaresi & Dressler, 2020)
- existing theories regarding the productivity–rivalry nexus, predicated upon non-evaluative rivalry, may not be applicable here, as rivalry within EM is controlled differently (Grandi, 2023) → how to expand quantitative approaches to explore also the pragmatics of EM?
- as the polyfunctionality of derivational processes, and consequently rivalry, is best evaluated through fine-grained semantic analysis (Rainer, 2014; Huyghe & Varvara, in press), developing more detailed annotations should be considered

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thank you!

ivan.lacic2@unibo.it

References I

Amiot, D. (2004).Haut degré et préfixation. *Travaux Linguistiques du Cerlico*, *17*, 91–104.

Baayen, H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 899–919). De Gruyter. DOI: 10.1515/9783110213881.2.899.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009).The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*, 209–226. DOI: 10.1007/s10579-009-9081-4.

Baroni, M., & Evert, S. (2014). The zipfR package for lexical statistics: A tutorial introduction [Available at: https://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf].

Bauer, L. (2001). *Morphological productivity*. Cambridge University Press. DOI: 10.1017/CBO9780511486210.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. DOI: 10.1017/9781316410899.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins. DOI: 10.1075/tsl.9.

Calpestrati, N. (2017). Intensification strategies in German and Italian written language. In M. Napoli & M. Ravetto (Eds.), *Exploring intensification: Synchronic, diachronic & cross-linguistic perspectives* (pp. 305–326). John Benjamins. DOI: 10.1075/slcs.189.16cal.

Cartier, E., & Huyghe, R. (2021).La concurrence affixale en diachronie: Le cas des préfixes de haut degré en français. *Linx*, *82*. DOI: 10.4000/linx.8078.

Dressler, W. U., Merlini Barbaresi, L., Schwaiger, S., Ransmayr, J., Sommer-Lolei, S., & Korecky-Kröll, K. (2019). Rivalry and lack of blocking among Italian and German diminutives in adult and child language. In F. Rainer, F. Gardani, W. U. Dressler, & H. C. Luschützky (Eds.), *Competition in inflection and word-formation* (pp. 123–143). Springer.

Evert, S., & Baroni, M. (2007).zipfR: Word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*, 29–32.

References II

Fernández-Domínguez, J. (2017). Methodological and procedural issues in the quantification of morphological competition. In J. Santana-Lario & S. Valera-Hernández (Eds.), *Competing patterns in english affixation* (pp. 67–117). Peter Lang.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* Second edition. John Wiley & Sons. DOI: 10.1002/0471445428.

Fradin, B., Dal, G., Grabar, N., Lignon, S., Namer, F., Tribout, D., & Zweigenbaum, P. (2008).Remarques sur l'usage des corpus en morphologie. *Langages*, *171*(3), 34–59. DOI: 10.3917/lang.171.0034.

Gaeta, L. (2010). Accrescitivo. In R. Simone (Ed.), *Enciclopedia dell'italiano* (pp. 14–16). Istituto della Enciclopedia Italiana.

Gaeta, L., & Ricca, D. (2015). Productivity. In P. O. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-formation: An international handbook of the languages of europe–vol. 2* (pp. 842–858). De Gruyter. DOI: 10.1515/9783110246278-003.

Goodman, L. A., & Kruskal, W. H. (1959).Measures of Association for Cross Classifications. II: Further Discussion and References. *Journal of the American Statistical Association*, *54*(285), 123–163. DOI: 10.1080/01621459.1959.10501503.

Grandi, N. (2023). Evaluative morphology in the romance languages. In *Oxford research encyclopedia of linguistics*. Oxford Univeristy Press. DOI: 10.1093/acrefore/9780199384655.013.684.

Gwet, K. L. (2008).Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. DOI: 10.1348/000711006X126600.

Huyghe, R., & Varvara, R. (in press).Semantic granularity in derivation. *Linguistics Vanguard*.

Iacobini, C. (2004). Prefissazione. In M. Gorssman & F. Rainer (Eds.), *La formazione delle parole in italiano* (pp. 97–163). De Gruyter.

Izert, M. (2012).Préfixes *extra-* et *supra-* comme intensificateurs de la propriété en français contemporain. *Kwartalnik Neofilologiczny*, *59*, 437–446.

References III

Kastovsky, D. (1986).The problem of productivity in word formation. *Linguistics*, *24*(3), 585–600. DOI: https://doi.org/10.1515/ling.1986.24.3.585.

Lindsay, M., & Aronoff, M. (2013). Natural selection in self-organizing morphological systems. In N. Hathout, F. Montermini, & J. Tseng (Eds.), *Morphology in toulouse. selected prodeedings of décembrettes 7* (pp. 133–153).

Mauri, C., Ballarè, S., Goria, E., Cerruti, M., & Suriano, F. (2019). KIParla corpus: a new resource for spoken Italian. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), *Clic-it 2019 – proceedings of the sixth italian conference on computational linguistics.*

Merlini Barbaresi, L., & Dressler, W. U. (2020). Pragmatic explanations in morphology. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Word knowledge and word usage: A cross-disciplinary guide to the mental lexicon* (Vol. 405–452). De Gruyter. DOI: 10.1515/9783110440577-011.

Montermini, F. (2008). *Il lato sinistro della morfologia: La prefissazione in italiano e nelle lingue del mondo.* FrancoAngeli.

Mutz, K. (2015). Evaluative morphology in a diachronic perspective. In N. Grandi & L. Körtvélyessy (Eds.), *Edinburgh handbook of evaluative morphology* (pp. 142–154). Edinburgh University Press. DOI: 10.1515/9780748681754-015.

Namer, F., Hathout, N., Amiot, D., Barque, L., Bonami, O., Boyé, G., Calderone, B., Cattini, J., Dal, G., et al. (2023).Démonette-2, a derivational database for French with broad lexical coverage and fine-grained morphological descriptions. *Lexique*, *33*, 6–40. DOI: ff10.54563/lexique.1242.

Napoli, M. (2012).Uno stra-prefisso: L'evoluzione di *stra-* nella storia dell'italiano. *Rivista italiana di linguistica e dialettologia*, *14*, 89–112.

Napoli, M., & Ravetto, M. (2017). Ways to intensify. Types of intensified meanings in Italian and German. In M. Napoli & M. Ravetto (Eds.), *Exploring intensification. Synchronic, diachronic and cross-linguistic perspectives* (pp. 327–352). John Benjamins. DOI: 10.1075/slcs.189.17nap.

Plag, I. (2003). *Word-formation in english.* Cambridge University Press. DOI: 10.1017/9781316771402.

References IV

Rainer, F. (2014). Polysemy in derivation. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of derivational morphology* (pp. 338–353). Oxford University Press. DOI: 10.1093/oxfordhb/9780199641642.013.0019.

Shannon, C. E. (1948).A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Silveira, P. S. P., & Siqueira, J. O. (2023).Better to be in agreement than in bad company: A critical analysis of many kappa-like tests. *Behavior Research Methods*, *55*(7), 3326–3347. DOI: https://doi.org/10.3758/s13428-022-01950-0.

Vach, W., & Gerke, O. (2023).Gwet's AC1 is not a substitute for Cohen's kappa–A comparison of basic properties. *MethodsX*, *10*, 102212. DOI: https://doi.org/10.1016/j.mex.2023.102212.

Ylonen, T. (2022). Wiktextract: Wiktionary as machine-readable structured data. In N. Calzolari et al. (Eds.), *Proceedings of the 13th conference on language resources and evaluation (lrec)* (pp. 1317–1325). URL: https://aclanthology.org/2022.lrec-1.140.

Zanchetta, E., & Baroni, M. (2005).Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings from the Corpus Linguistics Conference Series 2005.*

Zingarelli, N., Cannella, M., Lazzarini, B., & Zaninello, A. (2023). *Lo zingarelli 2024. vocabolario della lingua italiana.* Zanichelli.