Incidence- and abundance-based measures to assess rivalry in word formation

Justine Salvadori, Rossella Varvara, Richard Huyghe [firstname.lastname@unifr.ch]

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

1 Background

Affix rivalry occurs between affixes that have equivalent semantic functions and can compete in the formation of derivatives [1]. However, affixes are rarely strictly equivalent due to their **polyfunctionality**.

For example, suffixes used to form **agent nouns** in French also have unshared semantic functions.

	Agent	Instrument	Beneficiary	Inhabitant	Container	Partisan
-aire	✓	–	✓	–	✓	–
-ant	✓	✓	✓	–	–	✓
-eur	✓	✓	–	–	–	–
-ien	✓	–	–	✓	–	–
-ier	✓	✓	–	–	✓	–
-iste	✓	–	–	✓	–	✓

Table 1. Subset of semantic types realized by 6 polyfunctional suffixes in French

Different **degrees** of rivalry can be postulated depending on how (dis)similar affixes are. A **coefficient** of competition may be useful to compare situations of rivalry both within languages and cross-linguistically.

Objective: Explore measures of **semantic similarity** between polyfunctional affixes that can be used to assess their partial rivalry.

2 Similarity measures

We consider two measures drawn from studies in ecology. They both range from **0** (= **full dissimilarity**) to **1** (= **identity**).

1. The Sørensen index [2] is based on **presence/absence** data, depending on the number of distinct functions realized by rival affixes. It quantifies how similar two affixes are according to the **proportion** of functions they share.

$$S = \frac{2\,|A \cap B|}{|A| + |B|}$$

A = set of functions of Affix α
B = set of functions of Affix β
$A \cap B$ = set of functions common to α and β
$|X|$ = number of elements included in Set X

2. The Percentage similarity coefficient [3] is based on **abundance** data, depending on the number of derivatives realizing each function of rival affixes. It quantifies how similar two affixes are considering **type frequencies**.

$$PS = \frac{2\sum_{i=1}^{p} min(N_{i\alpha}, N_{i\beta})}{\sum_{i=1}^{p}(N_{i\alpha} + N_{i\beta})}$$

$N_{i\alpha}$ = number (i.e., the abundance) of derivatives with Affix α that realize Function i
$N_{i\beta}$ = number of derivatives with Affix β that realize Function i
p = total number of functions observed for α and β
$min(a,b)$ = the smaller of two numbers a and b

3 Fake data

The behavior of the Sørensen (*S*) index and the Percentage similarity (*PS*) coefficient can be illustrated with fake data.

		F1	F2	F3	F4	*S*	*PS*
Pair 1	Affix A	30	30	30	30	0.86	0.75
	Affix B	40	40	40	0		
Pair 2	Affix A	30	30	30	30	0.86	0.33
	Affix C	110	5	5	0		

S returns the same value for two pairs of rival affixes with the same functions, regardless of the frequency of realization of functions.

		F1	F2	F3	F4	*S*	*PS*
Pair 3	Affix D	20	20	20	0	0.67	0.67
	Affix E	20	20	0	20		
Pair 4	Affix F	10	10	20	20	1.00	0.67
	Affix G	20	20	10	10		

Tables 2-3. Number of derivatives per semantic function (F1-F4) and similarity scores (*S*, *PS*) obtained for pairs of rival affixes

PS returns the same value for two pairs of rival affixes with the same ratio between the minimal number of derivatives with shared functions and the total number of derivatives, regardless of the number of shared functions.

S and *PS* capture **different aspects** of similarity. Taken together, they provide a detailed description of similarity relationships.

4 Case study

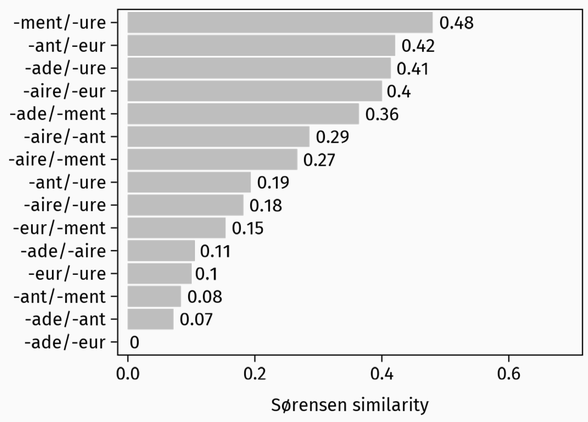
The potential of the two measures is explored using real linguistic material, viz. rival suffixes used to form **deverbal nouns in French**.

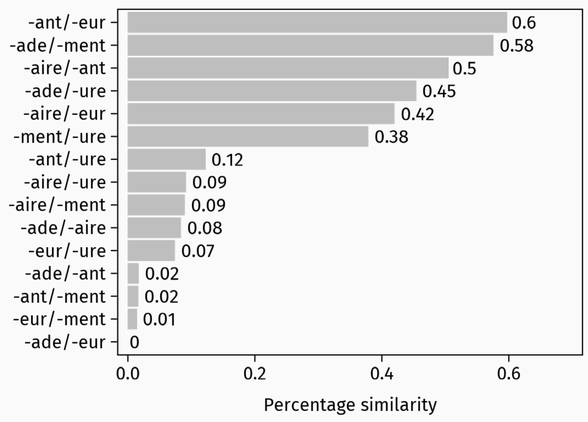
Data collection: 3 **eventive** (*-ade*, *-ment*, *-ure*) and 3 **agentive** suffixes (*-aire*, *-ant*, *-eur*) are selected. A random sample of **100** French deverbal nouns formed with each suffix is retrieved from the French web corpus FRCOW16A [4].

Semantic analysis: Each derived noun is analyzed using a **double classification** [5], distinguishing between the **ontological** description of the referent (e.g. event, animate, artefact) and the **relation** with the eventuality denoted by the base verb (e.g. agent, result, location).

Computation of scores: *S* and *PS* are applied to the suffixes based on the **782** word meanings/**37** functions identified in the dataset.

Results: The *S* (M = 0.23, SD = 0.15) and *PS* (M = 0.23, SD = 0.23) scores are presented in Figures 1-2.





Figures 1-2. Similarity scores. Pairs of suffixes are ordered from top to bottom by decreasing similarity

1. The diversity of the scores supports a **gradient** approach to affix rivalry. Almost all suffixes compete (even in very small proportions) and there are no perfect rivals in the sample.

2. The *S* and *PS* scores are **correlated** (Mantel test: r = .875, $p < .01$). Suffixes that have many functions in common also tend to present a relatively even distribution of derivatives across shared functions.

3. Suffixes belonging to the same **semantic group** ("eventive" or "agentive") are more similar to each other than to suffixes included in another group. This trend is more pronounced for *PS* than for *S* scores. The general distinction between the two types of suffixes relies more on the **frequent realization** of identical functions than on their ability to serve one function in particular.

5 Conclusion

Summary: This study introduces the Sørensen index (*S*) and the Percentage similarity coefficient (*PS*) as **measures of affix rivalry** and explores their potential through the analysis of a sample of 600 nouns formed with 6 nominalizing suffixes in French. The metrics highlight different aspects of **functional similarity** between affixes and should be considered a **first step** towards a comprehensive measurement of morphological competition.

Future work: *S* and *PS* could be examined **diachronically** over different periods of time and could also be combined with **productivity** measures in order to improve the assessment of rivalry.

References

[1] Huyghe, R., & Varvara, R. (2023). Affix rivalry: Theoretical and methodological challenges. *Word Structure*, *16*(1), 1-23.

[2] Sørensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.*, *5*, 1-34.

[3] Odum, E. P. (1950). Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology*, *31*(4), 587-605.

[4] Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the Web using a new efficient tool chain. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). European Language Resources Association.

[5] Salvadori, J., & Huyghe, R. (2023). Affix polyfunctionality in French deverbal nominalizations. *Morphology* 33, 1–39.

GitHub