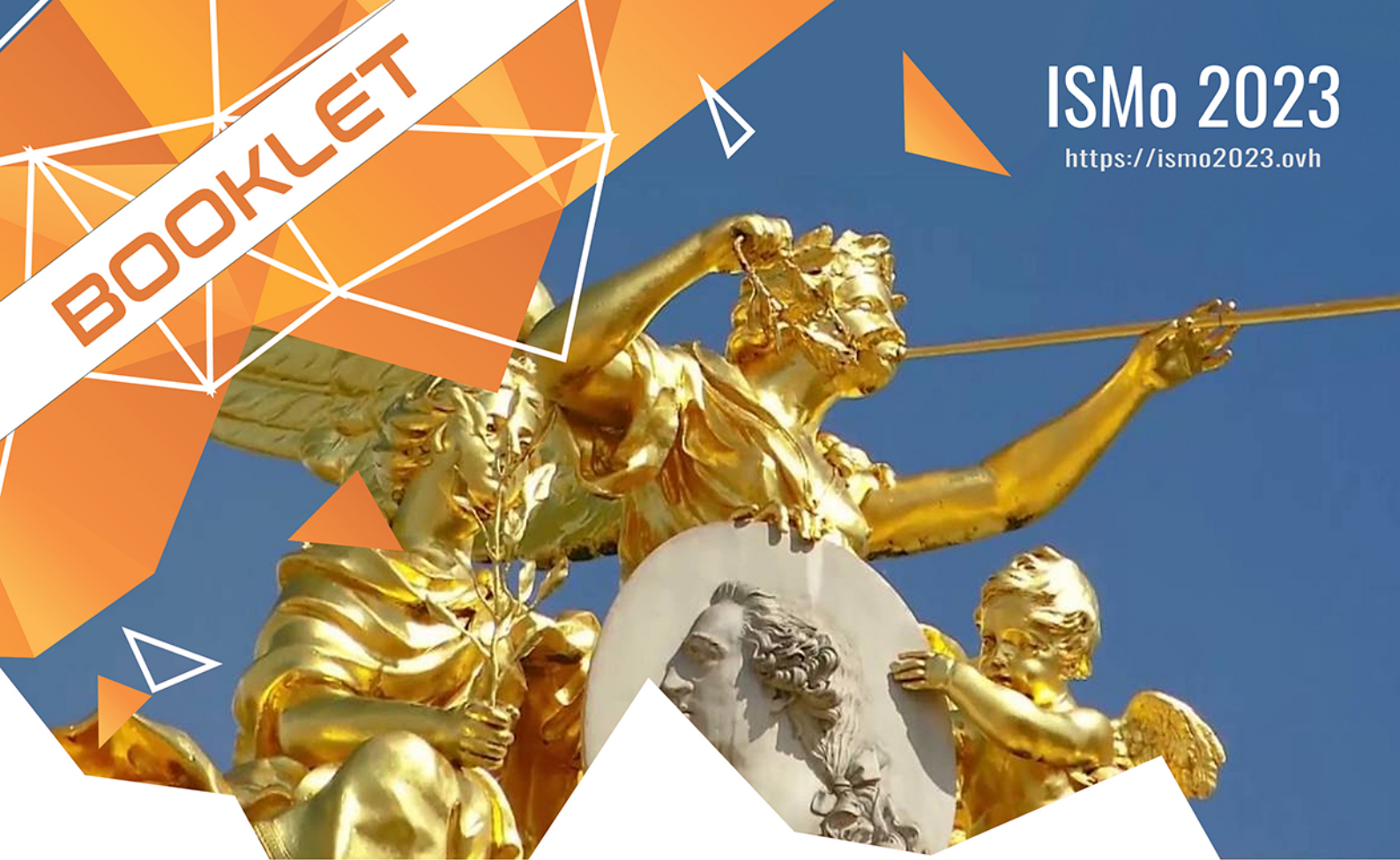


BOOKLET

ISM<sub>o</sub> 2023

<https://ismo2023.ovh>



# International Symposium of Morphology 2023

**GUEST SPEAKERS**

Susan Olsen (Humboldt-Universität zu Berlin)


Livio Gaeta (Università di Torino)

13 > 15  
september  
2023

**NANCY** | Campus Lettres &  
Sciences Humaines

**ATILF** | Building CNRS  
Room Paul Imbs, 2<sup>nd</sup> floor





**Fourth International Symposium of Morphology  
(ISM<sub>o</sub> 2023)**

**Fiammetta Namer & Stéphanie Lignon**

**13-15 September 2023**





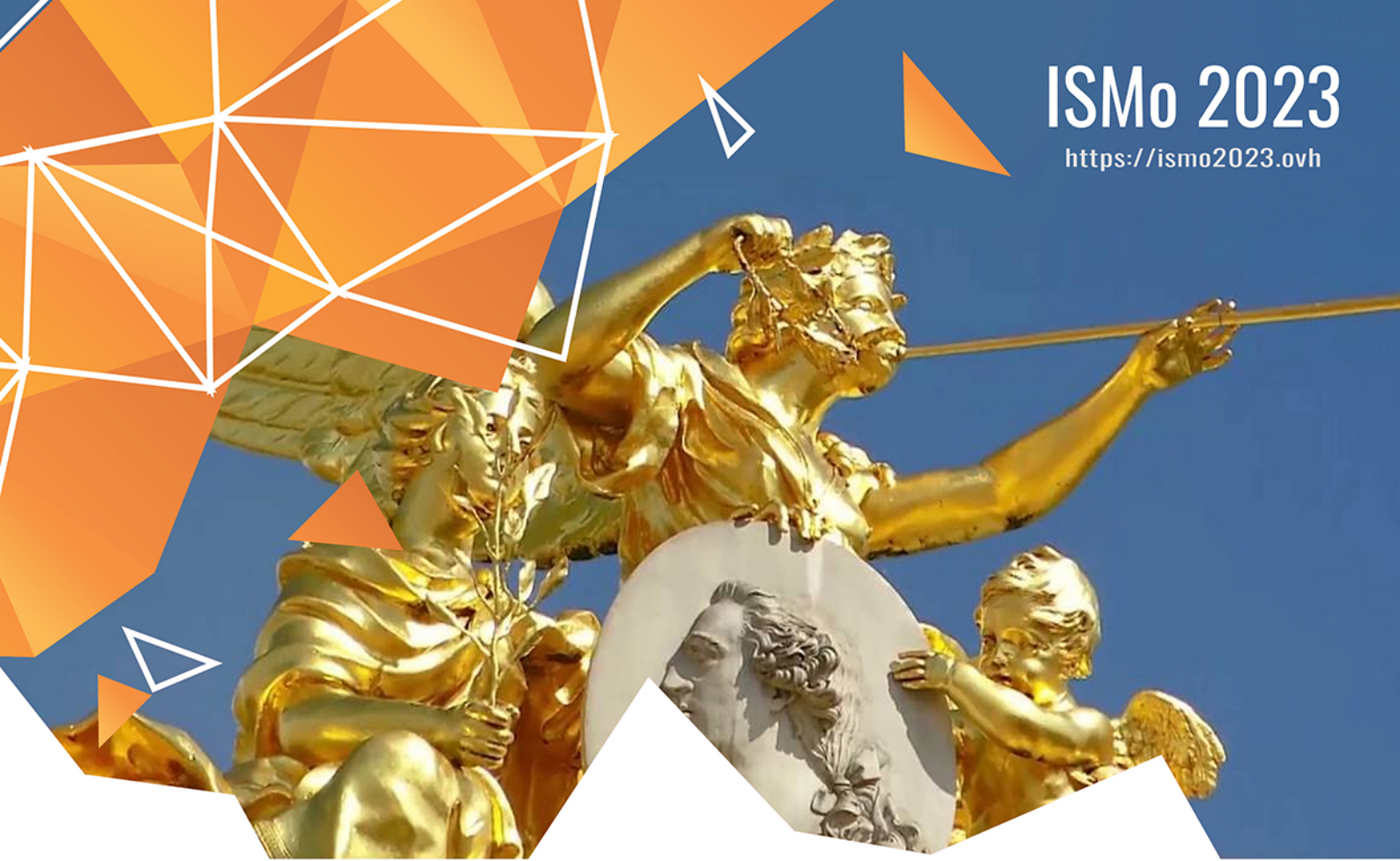
# CONTENT

<b>Content</b>	<b>5</b>
<b>I - Program</b>	<b>7</b>
<b>Schedule</b>	<b>9</b>
Wednesday, 13 <sup>th</sup> September 2023	<b>9</b>
Thursday, 14 <sup>th</sup> September 2023	<b>10</b>
Friday, 15 <sup>th</sup> September 2023	<b>11</b>
<b>Committees</b>	<b>13</b>
Program Committee	<b>13</b>
ISMo Standing Committee	<b>14</b>
Acknowledgements	<b>15</b>
<b>II - Papers</b>	<b>17</b>
<b>Keynote papers</b>	<b>19</b>
Susan Olsen: <i>Parasynthetic, Synthetic and Exocentric Compounds</i>	<b>21</b>
Livio Gaeta: <i>Constructions ex nihilo: conversion, backformation, secretion and spandrels</i>	<b>22</b>
<b>Oral Presentations</b>	<b>33</b>
Justine Salvadori, Rossella Varvara, Richard Huyghe: <i>Quantitative measures of affix rivalry</i>	<b>35</b>
Daniele Sanacore, Nabil Hathout, Fiammetta Namer: <i>A story-based approach to derivational paradigms</i>	<b>39</b>
Martha Booker Johnson, Micha Elsner, Andrea D. Sims: <i>High frequency derived words have low semantic transparency mostly only if they are polysemous</i>	<b>43</b>
Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, Tiago Tresoldi: <i>Paralex: a DeAR standard for rich lexicons of inflected forms.</i>	<b>47</b>
Berthold Crysmann, Baptiste Loreau Unger: <i>Disentangling morphomic splits in Limbu.</i>	<b>51</b>
Borja Herce: <i>From sound change to overabundance: the history of wa-/wu- and mba-/mbu- prefixal allomorphy in Central Pame (Otomanguean).</i>	<b>55</b>
Carlos Benavides: <i>Compounding in the Slot Structure Model</i>	<b>59</b>
Jan Radimský, M. Silvia Micheli: <i>Verbal-nexus and attributive-appositive N+N compounds in Italian A diachronic study</i>	<b>64</b>
Kalle Glauch: <i>The Interplay of Morpho-Phonology and Semantics in the Production of Conceptual Subject-Verb-Number Agreement in German</i>	<b>68</b>
Hiroki Koga: <i>Leveling among Patterns of Prosodic Structures of Paradigms for Affix Allomorphy</i>	<b>72</b>
Manuel Badal: <i>Obscuring morphomic patterns: some evidence from Catalan verbal inflection</i>	<b>76</b>

Lior Laks: <i>From competing patterns to competing structures: Verbal constructions based on loanwords in Hebrew</i>	81
Yagmur Ozturk, Izabella Thomas, Snejana Gadjeva: <i>Defining Semantic Categories for the Description of Turkish Nominal Morphemes</i>	85
Bernard Fradin: <i>On the polysemy of derivational exponents</i>	89
Sacha Beniamine, Mae Carroll: <i>The other perspective on exponence</i>	93
Grigorij Sibiljof: <i>The boundaries of person parameters of variation</i>	97
László Palágyi: <i>Back-formation, forward-formation, and cross-formation in the same construction. The case of Hungarian compound verbs</i>	101
Aurélie Héois: <i>Nouns in -ion and denominal verbs: can the output be explained? The case of English and French</i>	105
Ryohei Naya, Takashi Ishida: <i>On Imaginary English Dvandvas in Relational Adjectives</i>	109
Adèle Hénot-Mortier: <i>Morpho-semantics of the French diminutive suffix -et(te)</i>	113
Marie Huygevelde, Ridvan Kayirici, Olivier Bonami, Barbara Hemforth: <i>Affix rivalry in French demonyms: an experimental approach</i>	117
Milena Belosevic: <i>Creativity in name-based word formation: Evidence from the experimental study of personal name blends</i>	121
Maria Copot, Olivier Bonami: <i>Baseless derivation: the behavioural reality of derivational paradigms</i>	125
Fabio Montermini, Natalia Bobkova: <i>A lexico-paradigmatic analysis of Russian demonyms</i>	129
<b>Poster presentations</b>	<b>133</b>
Thomas Bertin, Nicole Bessière: <i>Sémantique des adjectifs dénominaux suffixés en -u et construits à partir d'un nom d'élément du corps</i>	135
Concha Castillo-Orihuela: <i>The binary vs. privative status of verbal inflectional morphology: The case of Germanic</i>	139
Hugo Dumoulin: <i>A contribution of discourse analysis to the morphology of nominalizations. Study of the use of nominalizations in the genre of the scientific activity report using a corpus linguistics approach based on the Démonette and Lexique 3 databases.</i>	143
Sophie Ellsäßer: <i>Paradigmatic structures, defectivity, and the specificity of referents</i>	147
Stefan Hartmann, Kristian Berg, Daniel Claeser: <i>Morphology and spelling variation: A case study on handwritten German</i>	150
Kentaro Koga: <i>Base ellipsis in coordinative constructions: the case of pré et post-X in French</i>	154
Irene Lami, M. Silvia Micheli, Jan Radimský, Joost Van De Weijer: <i>Gender agreement in Italian compounds with capo-</i>	158
Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, Giovanni Moretti: <i>PrinParLat: a resource of Latin principal parts</i>	162
Erich Round, Sacha Beniamine, Louise Esher: <i>The role of paradigm-external anchoring in simulating the emergence of inflection class systems</i>	166
Florence Villoing, Marie Laurence Knittel, Philippe Grea, Rafael Marin: <i>Innovative uses of French neological -ance nominalizations</i>	170
Jiahui Zhu: <i>On the analysis of the neological resultative construction suffixed with -iser and 化 [-huà] in contemporary media</i>	174

ISM<sub>o</sub> 2023

<https://ismo2023.ovh>



# Part I

# Program





# SCHEDULE

Wednesday, 13<sup>th</sup> September 2023

**12:30** Registration (ATILF, ground floor)

**14:15** Opening

## Session 1 – chair: Berthold Crysmann

**14:30** Justine Salvadori, Rossella Varvara, Richard Huyghe  
*Quantitative measures of affix rivalry*

**15:00** Daniele Sanacore, Nabil Hathout, Fiammetta Namer  
*A story-based approach to derivational paradigms*

**15:30** Martha Booker Johnson, Micha Elsner, Andrea D. Sims  
*High frequency derived words have low semantic transparency mostly only if they are polysemous*

**16:00** Coffee break (ATILF, 1<sup>st</sup> floor)

## Session 2 – chair: Jan Radimský

**16:30** Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, Tiago Tresoldi  
*Paralex: a DeAR standard for rich lexicons of inflected forms*

**17:00** Berthold Crysmann, Baptiste Loreau Unger  
*Disentangling morphomic splits in Limbu*

Thursday, 14<sup>th</sup> September 2023

**08:30** Registration (ATILF, ground floor)

**08:45** Welcome – coffee (ATILF, 1<sup>st</sup> floor)

**09:00** Keynote lecture 1 (chair: Fabio Montermini): Susan Olsen  
*Parasyntetic, Synthetic and Exocentric Compounds*

## Session 3 – chair: Lior Laks

**10:00** Carlos Benavides  
*Compounding in the Slot Structure Model*

**10:30** Jan Radimský, M. Silvia Micheli  
*Verbal-nexus and attributive-appositive N+N compounds in Italian A diachronic study*

**11:00** Coffee break (ATILF, 1<sup>st</sup> floor)

## Session 4 – chair: Sacha Beniamine

- 11:30** **Kalle Glauch**  
*The Interplay of Morpho-Phonology and Semantics in the Production of Conceptual Subject-Verb-Number Agreement in German*
- 12:00** **Hiroki Koga**  
*Leveling among Patterns of Prosodic Structures of Paradigms for Affix Allomorphy*
- 12:30** **Manuel Badal**  
*Obscuring morphomic patterns: some evidence from Catalan verbal inflection*
- 13:00** Lunch (buffet) (CLSH Campus, building A, 1<sup>st</sup> floor, Room A104)

## Poster Session (ATILF, 1<sup>st</sup> and 2<sup>d</sup> floors)

- 14:30** **Thomas Bertin, Nicole Bessière**  
*Sémantique des adjectifs dénominaux suffixés en -u et construits à partir d'un nom d'élément du corps*
- Concha Castillo-Orihuela**  
*The binary vs. privative status of verbal inflectional morphology: The case of Germanic*
- Hugo Dumoulin**  
*A contribution of discourse analysis to the morphology of nominalizations. Study of the use of nominalizations in the genre of the scientific activity report using a corpus linguistics approach based on the Démonette and Lexique 3 databases*
- Sophie Ellsäßer**  
*Paradigmatic structures, defectivity, and the specificity of referents*
- Kentaro Koga**  
*Base ellipsis in coordinative constructions: the case of pré et post-X in French*
- Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, Giovanni Moretti**  
*PrinParLat: a resource of Latin principal parts*
- Erich Round, Sacha Beniamine, Louise Esher**  
*The role of paradigm-external anchoring in simulating the emergence of inflection class systems*
- Florence Villoing, Marie Laurence Knittel, Philippe Grea, Rafael Marin**  
*Innovative uses of French neological -ance nominalizations*
- Jiahui Zhu**  
*On the analysis of the neological resultative construction suffixed with -iser and 化 [-huà] in contemporary media*
- 16:00** Coffee break (ATILF, 1<sup>st</sup> floor)

## Session 5 – chair: Nabil Hathout

- 16:30** **Lior Laks**  
*From competing patterns to competing structures: Verbal constructions based on loanwords in Hebrew*
- 17:00** **Yagmur Ozturk, Izabella Thomas, Snejana Gadjeva**  
*Defining Semantic Categories for the Description of Turkish Nominal Morphemes*
- 17:30** **Bernard Fradin**  
*On the polysemy of derivational exponents*
- 18:00** Welcome Drink (CLSH Campus, building A, 1<sup>st</sup> floor, Room A104)

# Friday 15<sup>th</sup> September 2023

**08:45** Welcome – coffee (ATILF, 1<sup>st</sup> floor)

**09:00** Keynote lecture 2 (chair: **Olivier Bonami**): **Livio Gaeta**  
*Constructions ex nihilo: conversion, backformation, secretion and spandrels*

## Session 6 – chair: Florence Villoing

**10:00** **Sacha Beniamine, Mae Carroll**  
*The other perspective on exponence*

**10:30** Coffee break (ATILF, 1<sup>st</sup> floor)

## Session 7 – chair: Livio Gaeta

**11:30** **László Palágyi**  
*Back-formation, forward-formation, and cross-formation in the same construction. The case of Hungarian compound verbs*

**12:00** **Aurélie Héois**  
*Nouns in -ion and denominal verbs: can the output be explained? The case of English and French*

**12:30** **Ryohei Naya, Takashi Ishida**  
*On Imaginary English Dvandvas in Relational Adjectives*

**13:00** Lunch (buffet) (CLSH Campus, building A, 1<sup>st</sup> floor, Room A104)

## Session 8 – chair : Bernard Fradin

**14:30** **Adèle Hénot-Mortier**  
*Morpho-semantics of the French diminutive suffix -et(te)*

**15:00** **Marie Huygevelde, Ridvan Kayirici, Olivier Bonami, Barbara Hemforth**  
*Affix rivalry in French demonyms: an experimental approach*

**15:30** **Milena Belosevic**  
*Creativity in name-based word formation: Evidence from the experimental study of personal name blends*

**16:00** Coffee break (ATILF, 1<sup>st</sup> floor)

## Session 9 – chair: Susan Olsen

**16:30** **Maria Copot, Olivier Bonami**  
*Baseless derivation: the behavioural reality of derivational paradigms*

**17:00** **Fabio Montermini, Natalia Bobkova**  
*A lexico-paradigmatic analysis of Russian demonyms*

**17:30** Closure



# COMMITTEES

## Program committee

We would like to express our deepest gratitude to the members of the program committee, for their expertise on the papers submitted to this conference.

- Fiammetta Namer & Stéphanie Lignon (ISM0 2023 co-chairs)
- Dany Amiot (Université de Lille)
- Giorgio Francesco Arcodia (Università di Milano-Bicocca)
- Mark Aronoff (Stony Brook University)
- Jenny Audring (Leiden University)
- Sacha Beniamine (Surrey Morphology Group)
- Olivier Bonami (Université de Paris-Cité)
- Gilles Boyé (Université Bordeaux-Montaigne)
- Dunstan Brown (University of York)
- Basilio Calderone (CNRS & Université de Toulouse Jean-Jaurès)
- Berthold Crysmann (CNRS & Université de Paris-Cité)
- Georgette Dal (Université de Lille)
- Serena Dal Maso (Università di Verona)
- Sebastian Fedden (LaCiTO, Sorbonne Nouvelle & CNRS)
- Bernard Fradin (LLF, CNRS & Université de Paris-Cité)
- H  l  ne Giraud (CLLE, CNRS & Universit   de Toulouse Jean-Jaur  s)
- Nabil Hathout (CLLE, CNRS & Universit   de Toulouse Jean-Jaur  s)
- Claudio Iacobini (Universit   di Salerno)
- Jean-Pierre Koenig (University at Buffalo, Buffalo, NY)
- St  phanie Lignon (Universit   de Lorraine)
- Claudia Marzi (ILC-CNR)
- Francesca Masini (Universit   di Bologna)
- Chiara Melloni (Universit   di Verona)
- Fabio Montermini (CLLE, CNRS & Universit   de Toulouse Jean-Jaur  s)
- Fiammetta Namer (Universit   de Lorraine)
- Ren  ta Panocov   (Pavol Jozef   af  rik University in Ko  ice)
- Vito Pirrelli (ILC-CNR)
- Jan Radimsk  y (University of South Bohemia in   esk   Bud  jovice)
- Franz Rainer (WU Vienna)

- Andrea Sims (Ohio State University)
- Andrew Spencer (University of Essex)
- Pavol Štekauer (Pavol Jozef Šafárik University in Košice)
- Pavel Štichauer (Charles University in Prague)
- Gregory Stump (University of Kentucky)
- Pius ten Hacken (Leopold-Franzens-Universität Innsbruck)
- Anna Maria Thornton (Università dell'Aquila)
- Juliette Thuilier (Université de Toulouse Jean-Jaurès)
- Delphine Tribout (Université de Lille)
- Florence Villoing (Université Paris Nanterre)
- Madeleine Voga (Université Montpellier III)
- Marine Wauquier (Université Sorbonne Nouvelle)

## **ISM0 Standing Committee**

- Dany Amiot (STL / Université de Lille)
- Olivier Bonami (LLF / Université Paris-Cité)
- Gilles Boyé (CLLE / Université de Bordeaux)
- Berthold Crysmann (LLF / Université Paris-Cité)
- Georgette Dal (STL / Université de Lille)
- Bernard Fradin (LLF / Université Paris-Cité)
- Hélène Giraud (CLLE / Université Toulouse Jean Jaurès)
- Nabil Hathout (CLLE / Université Toulouse Jean Jaurès)
- Stéphanie Lignon (ATILF / Université de Lorraine)
- Fabio Montermini (CLLE / Université Toulouse Jean Jaurès)
- Fiammetta Namer (ATILF / Université de Lorraine)
- Delphine Tribout (STL / Université de Lille)
- Florence Villoing (MoDYCO / Université Paris Nanterre)
- Marine Wauquier (Lattice / Université Sorbonne Nouvelle)

# ACKNOWLEDGEMENTS

ISMo 2023 is organised by the [ATILF](#) Research laboratory (UMR 7118, CNRS & Université de Lorraine).

We would like to thank our partner institutions:



[Analyse et Traitement Informatique de la Langue Française | ATILF](#)  
(UMR 7118, CNRS & Université de Lorraine)



[Cognition, Langues, Langage, Ergonomie | CLLE](#)  
(UMR 5263, CNRS & Université Toulouse Jean Jaurès)



[Centre National de la Recherche Scientifique | CNRS](#)



[Langues, Textes, Traitements informatiques, Cognition | Lattice](#)  
(UMR 8094, CNRS & École Normale Supérieure & Université Paris Cité)



[Laboratoire de Linguistique Formelle | LLF](#)  
(UMR 7110, CNRS et Université de Paris Cité)



[Modélisation, Dynamiques, Corpus | MoDyCo](#)  
(UMR 7114, CNRS & Université Paris Nanterre)



[Savoirs, Textes, Langages | STL](#)  
(UMR 8163, CNRS & Université de Lille)



[Université de Lorraine](#)  
[Pôle Scientifique Connaissance, Langage, Communication, Sociétés | CLCS](#)



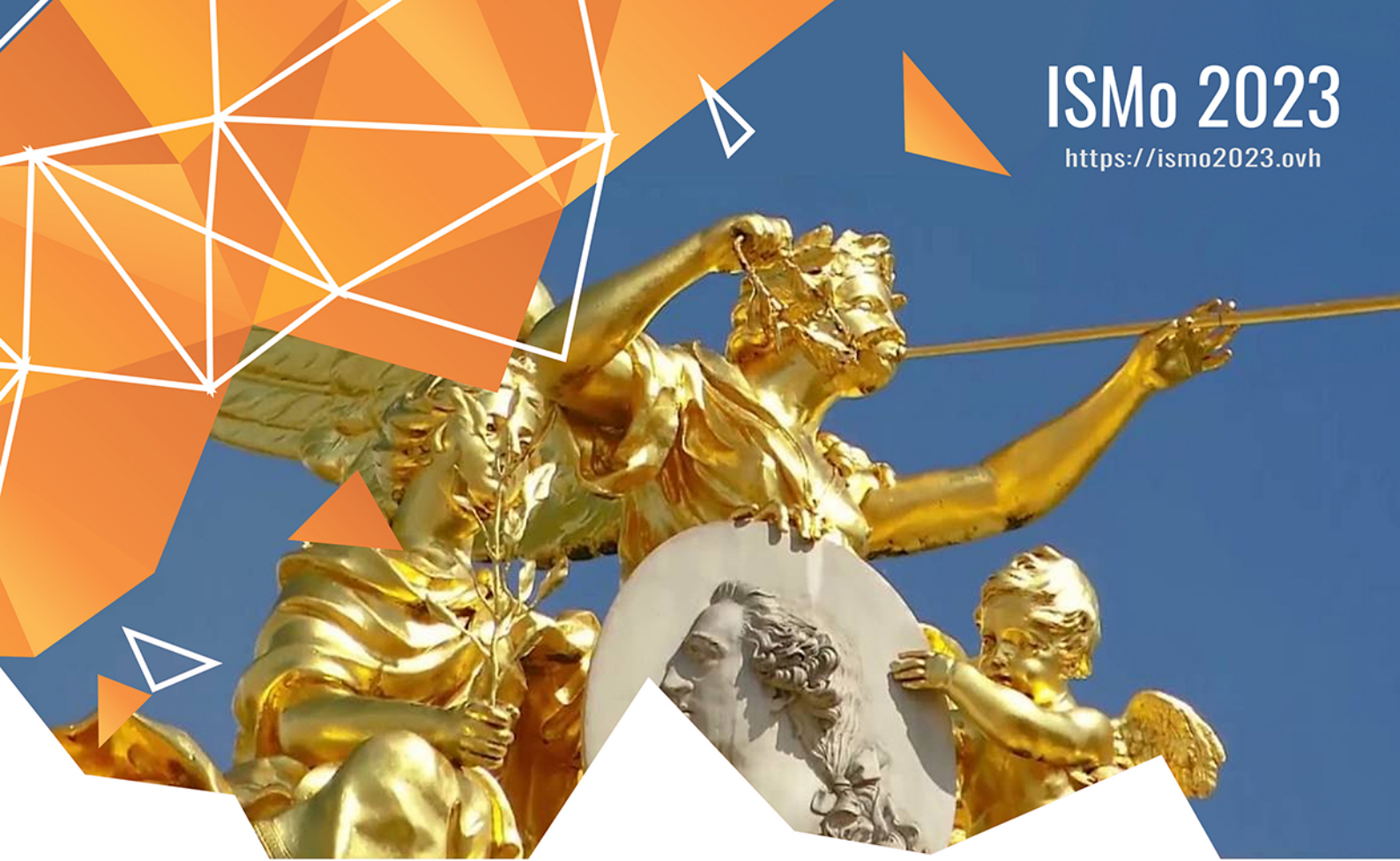
[Université de Lorraine](#)  
[UFR Sciences Humaines et Sociales, Nancy](#)





ISM<sub>o</sub> 2023

<https://ismo2023.ovh>



# Part II

# Papers



# **KEYNOTE PAPERS**



---

## Parasyntetic, Synthetic and Exocentric Compounds

*Susan Olsen*

Humboldt-Universität zu Berlin

---

Attempts to define the term *exocentric compound* and characterize the scope of exocentric linguistic configurations in the world's languages are manifold and result for the most part in different arrays of structural data. cf. Bauer (2008), (2010), Scalise & Guevara (2006), Scalise, Fábregas & Forza (2009), Ten Hacken (2010) among others. This lecture will approach the issue from a different angle, that is, not from a broad typological perspective, but first by focusing on the empirical data from one language family, Indo-European. The aim is to see whether, when studied in this manner, the notion *exocentric compound* has a coherent content and a useful function in the morphological description of the IE family. The discussion will attempt to reconstruct the origin of the term *exocentric compound* and compare it to the earlier terms *parasyntetic* and *synthetic* that have a longer tradition in the traditional discussion and in part overlap and in part contrast with the later notion of exocentric. It will ask what formations the term originally applied to, to determine when the term was first coined and aim to show how the range of data it originally encompassed has shifted in a uniform manner in the diachronic history of this larger IE family, leading to a slightly modified understanding of the term today than what was in its original focus.

### References

Bauer, Laurie, 2008. Exocentric compounds. *Morphology*, 18, 51-74.

Bauer, Laurie, 2010. The typology of exocentric compounding. In: Sergio Scalise and Irene Vogel (eds.), *Cross-disciplinary Issues in Compounding*, 167–176. Amsterdam: Benjamins.

Scalise, Sergio and Emiliano Guevara, 2006. Exocentric compounding in a typological framework. In: *Lingue e Linguaggio* vol.2, 185-206.

Scalise, Sergio, André Fábregas and F. Forza, F., 2009. Exocentricity in compounding. *Gengo Kenkyu* 135, 49-84.

ten Hacken, Pius, 2010. Synthetic and exocentric compounds in a parallel architecture. In: Susan Olsen (ed.). *New Impulses in Word-Formation*, 233–251. Hamburg: Buske.

---

# Constructions ex nihilo: conversion, backformation, secretion and spandrels

*Livio Gaeta*  
Department of Humanities  
University of Turin

---

## 1 Introduction

In Construction Morphology (cf. Booij 2010, Masini & Audring 2019) derivational relations are generally represented in hierarchical terms, in which the mechanism of Default Inheritance is largely exploited. This is concretely visualized by means of vertical relations connecting more general patterns or schemas lying high up in the Constructicon and more specific constructions placed down. Constructions ordered in such vertical relations are usually piled up in terms of specificity and/or reach of the schema, while these two criteria are not really elaborated in the current literature. In this paper, the attempt will be made at refining this view, especially focusing on cases in which new constructions are inferred ex nihilo by speakers, i.e. where on the basis of the formal substance occurring in a given schema a more general one is created which is formally less specific than the extant one, as is typical of conversions, backformations, secretions and spandrels.

## 2 Hierarchies and Default Inheritance

Construction Morphology generally represents derivational networks in hierarchical terms, crucially relying on the mechanism of Default Inheritance. A simple case is given by the following relational network in which an array of English words is connected in a multidimensional space (for typographic reasons this is flattened on a bidimensional picture, cf. Gaeta & Angster 2019):

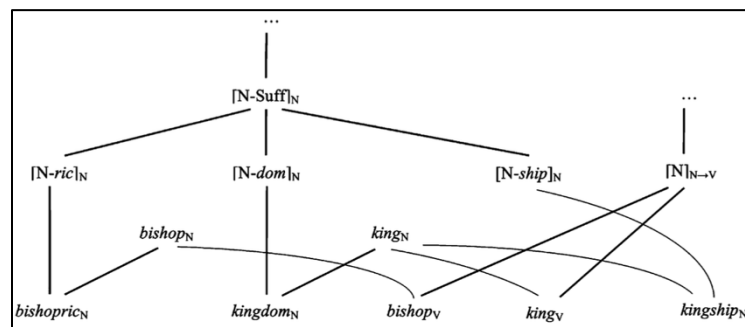


Fig. 1: Relational networking within the Constructicon

The vertical relations express increase in schematicity or generality, while the horizontal links represent the concrete connections holding among words entering specific morphological schemas. Accordingly, *bishop* is linked to *bishopric* and *to bishop*. The schemas are usually piled up in terms of specificity and/or reach of the schema. Thus, we observe that the schemas [N-ric]<sub>N</sub>, [N-dom]<sub>N</sub>, and [N-ship]<sub>N</sub>, are dominated by the more general schema [N-Suff]<sub>N</sub>, which is likely to be dominated by a less specific schema [X-Suff]<sub>X</sub>, and so forth. In these rough terms, however, nothing is said about the strength of the connections, both in the vertical and in horizontal dimension, namely about the productivity of the schemas. For instance, while *bishopric* is the only derivative formed with *-ric*, the conversion *to bishop* refers to a far richer pattern and forms a larger network of links, which is quantitatively heavier than that of *-ric*. Productivity is therefore likely to be expressed by the strength – in quantitative terms – of the

connections with which a certain schema is linked or networked. As is well known, this is not easy to operationalize, also because several possible understandings of productivity are currently suggested in the market. In previous research (cf. Gaeta 2016a), I defended the view that productivity must distinguish between the degree of entrenchment of a pattern which is expressed by its numerosity and by its expansive force expressed by the capacity of combining with new bases forming hapax legomena within a large text corpus. This distinction is important because it allows us to distinguish two possible types of networking. First, schemas which are fairly well entrenched, i.e. displaying a good degree of numerosity, but display a low expansive force, i.e. a scarce number of hapaxes. This is for instance the case of a suffix like Italian *-za* forming mostly deadjectival nouns (cf. *sapiente* ‘wise’ → *sapienza* ‘wisdom’) which presents a rich network of derivatives while its expansive force is reduced to sparse new formations. The opposite case is given by highly expansive evaluative prefixes like *mini-*, *mega-*, etc. which display a large number of hapaxes but are scarcely entrenched given that their network amounts to a dozen of lexemes. It is all but easy to incorporate this information into the light architecture of fig. 1. Nor is the impact of this information on the architecture easy to account for. For instance, one might wonder whether we are dealing with the same kind of phenomenon, i.e. the probability of applying a certain schema, or whether they are behaviourally distinct. Correspondingly, the first view of productivity appertains to the grammar, while the second one regards the discourse because evaluative markers are normally used in a certain type of utterances where they carry out a discourse-related function of a morpho-pragmatic sort. In the case of the suffix *-za*, on the other hand, nothing seems to hinge on a particular discourse situation for the derivative to be coined. To be true, there might be some bias of the text genre because of its possible employment in terminologies. But this does not seem to increase its expansive force in contrast to the morpho-pragmatic diffusion of evaluative prefixes. On the other hand, this view emphasizes that productivity has to be seen as a scalar notion, whose weight depends on the degree of networking of a schema. In the next sections, we will focus on this last aspect investigating the expansive force of schemas in networks.

### 3 Up and down the hierarchies

The syntagmatic and the paradigmatic dimension are clearly interwoven. This is evidenced in plastic form by cases in which a schema arises via the generalization of the criss-crossing of syntagmatic and paradigmatic relations (see Gaeta & Angster 2019). For instance, a German compound like *hochherzig* ‘generous-hearted’ can be treated as the result of telescoping two independent derivational mechanisms, a compositional one producing [Adj N]<sub>N</sub>-compounds (cf. *Rotwein* ‘red wine’, *Nationalstaat* ‘nation-state’, etc.) and a suffixal one [N-ig]<sub>Adj</sub> producing denominal adjectives (cf. *lustig* ‘funny’, *riesig* ‘huge’). As a result, two subschemas arise, a first one with *hoch-* ‘high-’ as modifier and a second one with *-herzig* ‘-hearted’ as head. In this latter, the telescoped interpretation obligatory refers to a metaphorical value of *Herz* ‘heart’. Therefore, *-herzig* selects adjectives compatible with this meaning, as shown in Fig. 2; on the other hand, the modifier *hoch-* narrows the possible filling of the N slot down to a range of nouns limited to those which are countable or allows for a scalar interpretation:

					...						
					<i>bös-</i>						
					<i>edel-</i>						
					<i>eng-</i>						
					<i>groß-</i>						

					<i>gut-</i>					
					<i>halb-</i>					
					<i>hart-</i>					
...	<i>-adlig</i>	<i>-fiebrig</i>	<i>-gradig</i>	<i>-hackig</i>	<b><i>hochherzig</i></b>	<i>-karätig</i>	<i>-klassig</i>	<i>-levelig</i>	<i>-oktanig</i>	...
					<i>kalt-</i>					
					<i>klein-</i>					
					<i>leicht-</i>					
					...					

Fig. 2: Criss-crossing of selective properties in German compounds

This can be plastically represented by means of a constructional network in which two different schemas are hierarchically subordinate and associated with their own selective properties:

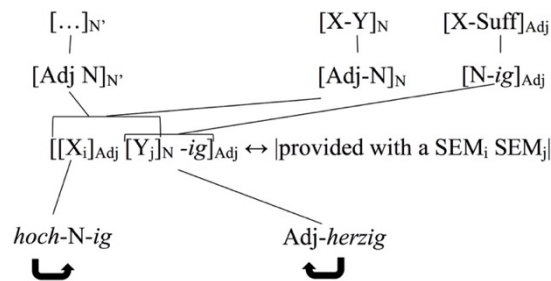


Fig. 3: Subschemas for German AN-ig compounds

In the different subschemas, the paradigmatic (selective) properties are driven by the different syntagmatic outlook. It is important to emphasize that the subschemas are independent of the occurrence of the intermediate derivational steps, such as for instance compounds like *\*Halbherz* or *\*Hochhack* or adjectives like *\*hackig* or *\*karätig*.

### 3.1 Telescoping and reanalysis

The effect of telescoping observed with *-herzig* is quite commonly found in several cases of reanalysis. For instance, in Spanish (see Rainer 1993: 483) a new suffix *-ería* is abducted on the basis of profession nouns ending with *-ero* to which a suffix *-ía* can be added forming nouns for the corresponding (work-)shop (1a); the new suffix is subsequently attached to simple nouns (1b):

(1)	a.	<i>cerveza</i> 'beer'	→	<i>cervecero</i> 'brewer'	→	<i>cervecería</i> 'brewery'
		<i>cristal</i> 'crystal'	→	<i>cristalero</i> 'glass dealer'	→	<i>cristalería</i> 'glassware shop'
		<i>joya</i> 'jewel'	→	<i>joyero</i> 'jeweler'	→	<i>joyería</i> 'jewelry store'
		<i>leche</i> 'milk'	→	<i>lechero</i> 'milkman'	→	<i>lechería</i> 'diary'
	b.	<i>acero</i> 'steel'	→		→	<i>acerería</i> 'steel mill'
		<i>estuco</i> 'plaster'	→		→	<i>estuquería</i> 'plaster workshop'
		<i>juguete</i> 'toy'	→		→	<i>juguetería</i> 'toy shop'
		<i>hamburguesa</i> 'burger'	→		→	<i>hamburguesería</i> 'burger point'



In a manner which is clearly similar to the German compounds with *-herzig*, we have here a process of multiple re-composition and pattern-expansion:

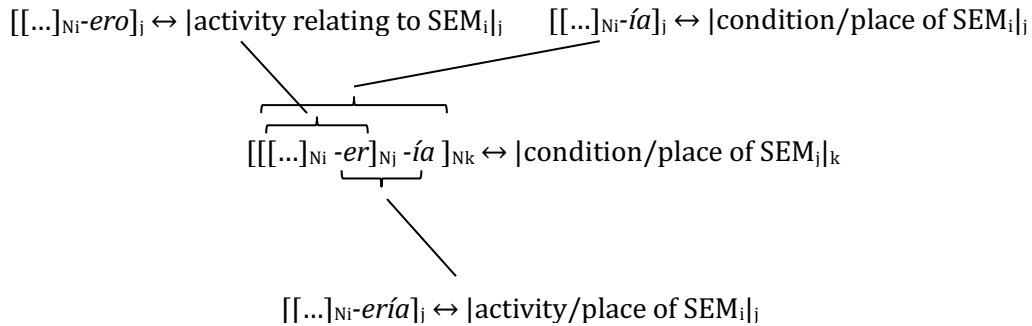


Fig. 2: Suffix telescoping for Spanish *-ería*

The re-composition is bolstered by the straightforward reanalysis of [[[*crystal*]-*er*]-*ía*] as [[*crystal*]-*ería*] simply dropping an intermediate step but keeping trace of its semantics because *cristalería* does not merely mean ‘store for crystal’ but ‘glassware shop’, i.e. a place where products resulting from the activity of a glass dealer or maker are sold. In this way we can account for the similar meaning of *estuqu-ería* ‘plaster workshop’, independently of the lack of \**estuquero*. The new suffix *-ería* goes well beyond the domain of the suffix *-ía* which is limited to nouns denoting human beings and belonging to well-curtailed domains like for instance church, administration, politics, as for instance *abad* ‘abbot’ → *abadía* ‘abbey’, *alcalde* ‘mayor’ → *alcaldía* ‘mayor’s hall / office’, *tirano* ‘tyrant’ → *tiranía* ‘tyranny’ and the like (cf. Rainer 1993: 512).

### 3.2 Secretion: constructions ex nihilo

The mirror-image case is given by the so-called secretion (Haspelmath 1995) in which a new affix arises by reusing segmental material which does not – or only partially – display a morphological status. In this sense, a new construction is created ex nihilo although some dust pre-exists which provides the concrete matter for the new entity. One clear example is given by the Italian word *grigiolino* ‘greyish’ which goes back to the simulative compound *grigio lino* ‘linen grey’, as shown by other similar colour compounds like *giallo limone* ‘lemon yellow’, *verde bottiglia* ‘bottle green’, *blu mare* ‘sea blue’, etc. Due to a false segmentation based on *magro* ‘slim’ → *magr-ol-ino* ‘slim-ol-DIM’, *scemo* ‘idiot’ → *scem-ol-ino* ‘idiot-ol-DIM’, *occhio* ‘eye’ → *occhi-ol-ino* ‘eye-ol-DIM’, *pesce* ‘fish’ → *pesci-ol-ino* ‘fish-ol-DIM’, a suffix *-ol-ino* has been extracted and extended to adjectives like *verde* ‘green’ → *verd-olino* ‘greenish’ and *beige* ‘beige’ → *beigi-olino* ‘beige-DIM’. Note that in *-ol-ino* the diminutive suffix *-ino* is expanded by means of an interfix which appears in a number of derivatives, typically in nouns as briefly exemplified above. In this way, two distinct types of diminutives based on *-ino* arise: *giall-ino*, *grig-ino*, *verd-ino*, *beig-ino*, and the new pattern *grigiolino*, *verdolino* and *beigiolino* (and *biancolino* ‘whitish’, *giallolino* ‘yellowish’ or *giallorino* with a phonological dissimilation, attested in old and modern texts):

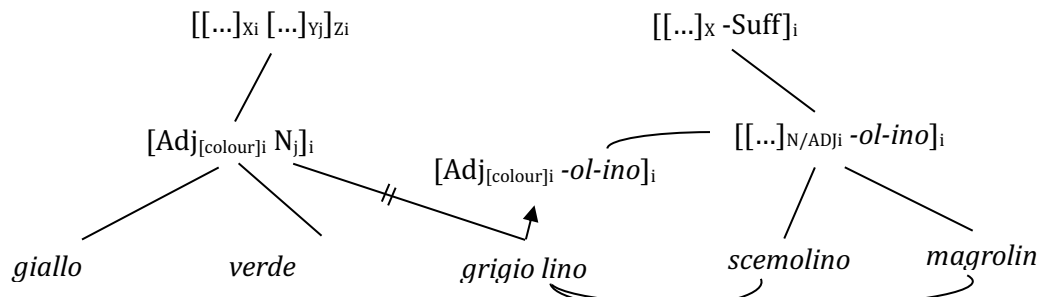


Fig. 3: The secretion of *grigiolino*

In this case, we observe the rise of a new subschema resulting from the general compound schema [Adj N] producing a certain number of formations. Because of a mere phonological similarity *grigiolino* is attracted by the series of derivatives formed with *-olino*. As a consequence, a new subschema for suffixation arises which partially inherits the properties of the compound schema specific of colour adjectives. At the same time, the link of *grigiolino* with the latter schema disappears, which corresponds to the loss of any activation of the lexeme *lino* and to the association with the series of *scemolino*, etc.

Besides the clear rhyming attraction exerted by the ending *-olino*, one additional factor which has played a role in the reanalysis is the peculiar status of alterative suffixes in Italian combined with the property of left-headedness of compounds. Italian alterative suffixes do not normally behave as heads as shown by common examples like *casa* ‘house[F]’ → *cas-etta* ‘house-DIM.F.’ vs. *libro* ‘book[M]’ → *libr-etto* ‘book-DIM.M.’ where the respective word properties (gender, inflectional class) of the diminutives filter directly from their bases. At the same time, Italian [Adj-N] compounds of this type are left-headed (in contrast to right-headed compounds like *auto-munito* ‘car-equipped’). On this basis, the switch of *grigiolino* from the compound to the suffixation schema does not change the crucial role of *grigio* as head. Furthermore, the case of *grigiolino* shows us that the two parallel domains of word-formation of compounding and affixation are reciprocally permeable. This is not only typical of grammaticalization of words into affixes, as repeatedly shown in the literature (cf. Gaeta 2022 for a brief discussion). Examples like the new subschema of *-olino* show that – besides and independently of grammaticalization – within morphology it is largely possible to re-assemble pieces of words exploiting the locally occurring information (in our case: the restriction to colour terms) and to give rise to new patterns which are “ergonomic” with regard to their systemic value, because the new pattern forms alteratives specifically for colour terms like those produced by simulative compounds.

While *-olino* might be considered marginal – a kind of spandrel, as we will see below – an example of secretion which has given rise to a highly productive pattern is given by the German suffix *-ler*, extremely productive with nominal bases for denoting agents, professions, etc.: *Kunst* ‘art’ → *Künstler* ‘artist’, *Sport* ‘sport’ → *Sportler* ‘sportsman’, *Schwergewicht* ‘heavyweight’ → *Schwergewichtler* ‘heavyweight boxer’, etc. The actual denominal suffix comes from the reanalysis of the suffix *-er* when employed with verbal bases ending with *-el-* like *angeln* ‘to angle’ / *Angler* ‘angler’, *nörgeln* ‘to grumble’ / *Nörgler* ‘grumbler’, *betteln* ‘to beg’ / *Bettler* ‘beggar’, etc. Due to the opacity resulting from the deletion of the unstressed *-e-*, a new suffix *-ler* was extracted and productively used specifically with nominal bases while *-er* is predominantly employed with verbs (cf. Eisenberg 2020: 440).

### 3.3 *Horror vacui*: conversion and backformation

In the preceding sections we have illustrated cases of “re-verticalization” of horizontal, syntagmatic relations which result from parallel and competing interpretations rising in the criss-cross of the lexical network. In particular, while telescoping refers to the shortening of a derivational sequence, its conceptual counterpart is backformation which in a way restores a transparent sequence filling up a real or potential hole in a network. Examples are dozens and generally result from the derivational processes complexifying words in a syntagmatic dimension. To be more explicit with concrete examples, I will mention two slightly different cases. The first one is given by the Italian verb like *decontribuire* ‘to de-contribute’ as it is found in the following examples (from the Internet):

- (2) a. *l'accordo del 23 luglio ... decontribuisce gli straordinari per meglio sfruttare i lavoratori*  
 ‘the July 23 agreement ... de-contributes overtime to better exploit workers’

- b. *dalla legge, che detassa e decontribuisce totalmente il salario corrisposto*  
 'by the law which totally de-taxes and de-contributes the salary paid'

While this is apparently a prefixation of *contribuire* 'to contribute' by means of the prefix *de-* which is productively found to form new verbs – see also in (2b) *detassare* 'to de-tax' – its meaning is completely odd. The reversative prefix *de-* normally serves as a modifier without any impact on the general meaning of the verbal base, including its argument structure (Iacobini 2004: 146): *comprimere* 'to compress' / *de-comprimere* 'to decompress', *congelare* 'to freeze' / *de-congelare* 'to defrost', *stabilizzare* 'to stabilize' / *de-stabilizzare* 'to destabilize', etc. Generally, the transitive base preserves its transitive argument structure, which constitutes one of the main pieces of evidence in support of the idea that they are not heads in contrast to most suffixes (Iacobini 2004: 106).

In the case of *de-contribuire* the intransitive base *contribuire* 'to contribute' undergoes prefixation which gives rise to a transitive verb with the special meaning: 'to reduce the social security charges paid on the salary'. Notice that the base *contribuire* does not normally convey the alleged meaning 'to pay security charges on the salary', but only an uncommon value 'pay taxes', within the general intransitive usage of the verb. The meaning inferable from the examples in (2) is clearly backderived from *decontribuzione*, already reported in dictionaries with the meaning 'reduction of the social security charges'. This results from the prefixation of the abstract noun *contribuzione* coined with the special meaning 'social security charges weighing on the gross salary'. Notice that *contribuzione* neatly contrasts with the common deverbal abstract *contributo* 'ACT/RESULT of contributing'. The account in terms of backformation explains not only the source of the specific meaning of *decontribuire* with regard to its unprefixated base, but also the apparent violation of the non-head value of the prefix which generally does not have an impact on the meaning of the verbal base and on its argument structure. The transitive value of the backderived verb *decontribuire* is likely to result from the reverbalization of nominal phrases reflecting the general pattern of abstract nouns in Italian as in *pagamento* 'payment' ← *pagare* 'to pay':

- (3) a. *Modalità di pagamento delle tasse*  
 'Methods of payment of fees'  
 b. *un provvedimento per la decontribuzione del salario*  
 'a provision for the decontribution of wages'  
 c. *insieme alla detassazione e decontribuzione degli straordinari*  
 'together with the detaxation and de-contribution of overtime'

In a parallel way with regard to the transitive bases *pagare* and *detassare*, a transitive verb *decontribuire* is backderived from the syntactic environment provided by the nominal phrase and reflected in the examples in (2) above. The account in terms of backformation, whereby *decontribuzione* based on the prefixation of *contribuzione* gives rise to *decontribuire*, apparently violates the restriction displayed by the reversative prefix *de-* that is only expected to combine with verbs and not with nouns (cf. Iacobini 2004: 112). However, this apparent violation is strongly supported by the occurrence of so-called unified or amalgamated schemas (Booij 2010) which result from the combination of simple schemas into more complex constructions based on highly recurrent patterns. Accordingly, we record both prefixed verbs lacking a verbal base resulting from conversion or suffixation (cf. *\*caffeinare*, *\*nuclearizzare* (4a-b)) and prefixed nouns in which the verbal base is either not attested or arguably backderived from the corresponding nouns (cf. *deprogrammare* 'cancel from the programming', but see *programmare*

'to plan, programme', *depenalizzare* 'to decriminalise', but see *penalizzare* 'to penalise, hinder' (4c-d), see Iacobini 2004: 147):

- (4) a. *decaffeinare* 'to decaffeinate'  
 $[de- [V]_V + [N]_V = [de- [N]_V]_V$   
 b. *denuclearizzare* 'to denuclearize'  
 $[de- [V]_V + [N-izza-]_V = [de- [N-izza-]_V]_V$   
 c. *deprogrammazione* 'de-programming'  
 $[de- [V]_V + [[N]_V -zione]_N = [de- [[N]_V -zione]_N$   
 d. *depenalizzazione* 'de-criminalisation'  
 $[de- [V]_V + [[N-izza-]_V -zione]_N = [de- [[N-izza-]_V -zione]_N$

On the other hand, backformation is also enhanced by so-called second-order schemas in which a paradigmatic relation is established between two distinct word-formation patterns. In our case, verbs formed with the suffix *-izzare* generally display a corresponding action noun in *-zione* as shown in (4d) above (see Gaeta 2004: 331):

- |  |  |
|--|--|
| (5) <i>realizzare</i> 'to realise'   | <i>realizzazione</i> 'realisation'           |
| <i>massimizzare</i> 'to maximise'  | <i>massimizzazione</i> 'maximisation'        |
| <i>inizializzare</i> 'to initalize'  | <i>inizializzazione</i> 'initialisation'     |
| <i>berlusconizzare</i> 'to berlusconise'   | <i>berlusconizzazione</i> 'berlusconisation' |
| $\langle [X-izza-]_V \leftrightarrow SEM_i \rangle \approx \langle [V-zione]_{N_j} \leftrightarrow [ACT\ of\ SEM_i]_j \rangle$ |  |

Furthermore, we record also a second-order schema for verbs prefixed with *de-* and the corresponding abstract noun suffixed with *-zione* as shown in (4c-d) above (see Gaeta 2004: 330):

- |   |   |
|---|---|
| (6) <i>deportare</i> 'to deport'  | <i>decapsulazione</i> 'deportation'         |
| <i>denudare</i> 'to undress'  | <i>denudazione</i> 'denudation'             |
| <i>declassare</i> 'to downgrade'  | <i>declassazione</i> 'downgrading'          |
| <i>desecretare</i> 'declassify'   | <i>desecretazione</i> 'to declassification' |
| $\langle [de-X]_{V_i} \leftrightarrow SEM_i \rangle \approx \langle [V-zione]_{N_j} \leftrightarrow [ACT\ of\ SEM_i]_j \rangle$ |   |

This network of more complex schemas based on other simpler schemas conspires in favouring a mechanism of backformation of the type described above for *decontribuire* providing paradigmatic support for possible syntagmatic gaps and restrictions:

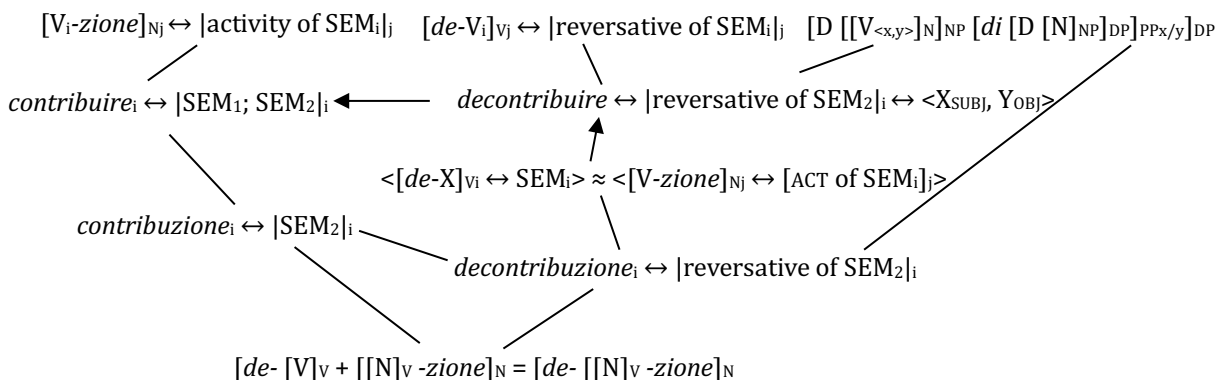


Fig. 4: The backformation of *decontribuire*

Notice in particular in the clearly simplified network summarized in Fig. 4 the role of the syntactic schema typical of noun phrases headed by a deverbal noun which is responsible for the argument structure of *decontribuire*, indirectly extracted from the syntactic behaviour of *decontribuzione* (see in this regard Gaeta & Zeldes 2017).

A second example elaborates on the concept of reverbalization which has been called into play for the explanation of the transitive argument structure of *decontribuire* with regard to the intransitive *contribuire*. This concept is in a way antiphrastic of the term conversion, in the sense that it represents the result of backformation with cases in which the simple verbal base already occurs in the lexicon as for *decontribuire*. A paramount example of reverbalization is given by German verbal compounds which generally result from the reverbalization of noun-based compounds (see Wurzel 1998, Gaeta 2014 for details). In particular, we can distinguish three types:

- |        |                                     |   |
|--------|-------------------------------------|---|
| (7) a. | <i>ehescheiden</i> ‘to divorce’     | <i>Ehe scheiden</i> ‘to divorce marriage’ / <i>Ehescheidung</i> |
| b.     | <i>eislaufen</i> ‘to ice-skate’     | <i>Eislauf</i> ‘ice skating’                                    |
| c.     | <i>bergsteigen</i> ‘to mountaineer’ | <i>Bergsteiger</i> ‘mountaineer’                                |

The first type in (7a) results from the reverbalization of a compound headed by the abstract noun *Scheidung* ‘separation, divorce’ which can also be treated straightforwardly as an incorporation of a direct object. In (7b) the reverbalization starts from a compound headed by the conversion *Lauf* ‘run’, while the reverbalization in (7c) is a true backformation in the sense that the compound is headed by the suffixed noun *Steiger* ‘head miner, riser’ reverbalized into the verb *steigen* meaning ‘to rise, climb’. In these last two types no account in terms of incorporation is available insofar as both verbs are intransitive. Reverbalization provides a unified mechanism to account for all these cases in which we clearly observe the interaction of the syntagmatic and paradigmatic properties of the constructions involved into the complex network of relations providing a productive way to form verbal compounds mediated by the support of unified and second-order schemas. As an extreme consequence, in German this mechanism can also give rise to denominal adjectives or adverbs like *klasse* or *spitze*, as they are found in the following expressions:

- |        |   |
|--------|---|
| (8) a. | <i>Bayern München hat klasse gespielt.</i><br>‘Bayern Munich played great (lit. class)’.                        |
| b.     | <i>Im Restaurant haben wir spitze gegessen.</i><br>‘In the restaurant we had a great meal (lit. we ate point)’. |

These adverbial usages come from the reverbalization of nominal compounds in which the modifier has an evaluative value like *Klassefrau* ‘great woman’, *Spitzenqualität* ‘prime quality’, etc. Since the head can also consist of a denominal abstract like for instance in *Klassenspiel* ‘great game’ or *Spitzenessen* ‘top food’, a reverbalization may occur giving rise to verbal compounds *klassemspielen* or *spitzenessen*. The latter share the property of separability which is typical of the German verbal compounds as shown by the following example:

- |     |   |
|-----|---|
| (9) | <i>Marta läuft eis und spielt klavier.</i><br>‘Marta ice-skates and plays the piano’. |
|-----|---|

By virtue of the property of separability, the mechanism of reverbalization creates the conditions for the adverbial usage of *klasse* and *spitze* as shown in (8) above.

### 3.4 Spandrels and exaptation

Spandrels are claimed by Gould & Lewontin (1979) to provide instances of pure recycling of biological material to serve a different purpose because they are intended as the by-product of the evolution of some other characteristic rather than a direct product of adaptive selection relating to (the functionality of) a certain organ. Thus, spandrels are peculiar instances of exaptation which do not display a pre-adaptive character in the sense that their properties and structure does not foreshadow any predisposition for the subsequent reuse in the new function. Adopting this perspective of exaptation and exaptive changes in linguistics (cf. Lass 1997: 316 and Gaeta 2016b for a general view), we can figure out cases of ‘spandrels’, namely of re-functionalisation of linguistic material occurring in a certain network of constructions which is pure ‘bricolage’ in the sense that the exapted form cannot be seen as pre-adapted for the new function as we have seen for the complex instance of backformation of *decontribuire* which is crucially supported by the occurrence of unified and second-order schemas. Furthermore, this also stands in neat contrast with the cases of reanalysis seen above like the telescoping of Spanish *-ería* and the secretion of German *-ler*. On the other hand, spandrels share the same drive of the above cases because they represent “re-verticalizations” aiming at escaping the *horror vacui*, namely at establishing morphological complexity even when this is not present etymologically. Examples of spandrels are indeed dozens. Besides classical examples like *-gate* and *-burger* extracted from the models *Watergate* and *hamburger*, where an alleged modifier *Water-* and *ham-* has been stripped away – also with the help of folk etymology (see Maiden 2020) – and replaced by a nest of words belonging to well-defined lexical sets, respectively [political scandals] and [sandwich fillers], we can mention several cases sharing the same jocular character which is in tune with Lass’ (1997: 309) view of the bricolage nature of exaptation. In spite of the traditional view that opacity might trigger the folk-etymological reanalysis, the most striking feature of at least some of these jocular spandrels is the completely transparent structure of the words serving as models. For instance, a spandrel *-elfie* has been extracted from the derivative *selfie*, although the latter is transparently derived from *self* with the addition of the diminutive suffix *-y/-ie*: *shelfie* ‘selfie made in front of a shelf’, *felfie* ‘farm animal selfie’, *lelfie* ‘selfie of legs’, *belfie* ‘selfie with bare buttocks’, *nelfie* ‘selfie of a nude person’, etc. (Hamans 2020).

## 4 Conclusion

To sum up, adopting a network approach to morphological complexity allows us to get rid of fallacies coming from top/down models like those characterizing structuralist linguistics or rule/list dichotomies typical of generative linguistics. Speakers constantly exploit the possibilities offered by the multiple links existing among words, including the complex morphosyntactic constructions in which they are involved, as shown by their influence on the argument structure of the backformation *decontribuire*. Re-verticalizations driven by *horror vacui* aim at exhausting the paradigmatic space and at restoring in this way transparency. On the other hand, spandrels are creatively produced which completely disregard the transparent complexity of the words involved. To be sure, we are a long way from discovering the limits of interpretation. On the other hand, a relational approach, crucially based on rich networks of more or less schematic constructions, is a promising method to pursue that goal.

## References

Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.

- Eisenberg, Peter. 2020. *Grundriss der deutschen Grammatik: Das Wort*. 5th ed. Berlin: Metzler.
- Gaeta, Livio. 2004. Nomi d'azione. In Maria Grossmann & Franz Rainer (eds.), *La formazione delle parole in italiano*, 314–351. Tübingen, Niemeyer.
- Gaeta, Livio. 2014. On decategorization and its relevance in German. In Raffaele Simone & Francesca Masini (eds.), *Word classes: nature, typology and representations*, 227–241. Amsterdam & Philadelphia: John Benjamins.
- Gaeta, Livio. 2016a. How lexical is morphology? The constructicon and the quadripartite architecture of grammar. In Livia Körtvélyessy, Pavol Štekauer & Salvador Valera (eds.), *Word-Formation across Languages*, 109–146. Cambridge: Cambridge Scholars Publishing.
- Gaeta, Livio. 2016b. Co-opting exaptation in a theory of language change. In Muriel Norde & Freek Van de Velde (eds.), *Exaptation in language change*, 57–92. Amsterdam & Philadelphia: John Benjamins.
- Gaeta, Livio. 2022. Dangerous Liaisons: An introduction to derivational paradigms. In Alba E. Ruz, Cristina Fernández-Alcaina & Cristina Lara-Clares (eds.), *Paradigms in Word Formation: Theory and Applications*, 3–18. Amsterdam & Philadelphia: John Benjamins.
- Gaeta, Livio & Marco Angster. 2019. Stripping paradigmatic relations out of the syntax. *Morphology* 29(2): 249–270.
- Gaeta, Livio & Amir Zeldes. 2017. Between VP and NN: On the Constructional Types of German -er Compounds. *Constructions & Frames* 9(1): 1–40.
- Gould, Stephen J. & Richard C. Lewontin. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B* 205: 581–598.
- Hamans, Camiel. 2020. Contra de linguïstische preutsheid: Over -gate en andere libfixen. *Nederlandse taalkunde* 25(2): 319–332.
- Haspelmath, Martin. 1995. The Growth of Affixes in Morphological Reanalysis. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1994*, 1–29. Dordrecht: Kluwer.
- Iacobini, Claudio. 2004. Prefissazione. In Maria Grossmann & Franz Rainer (eds.) *La formazione delle parole in italiano*, 97–164 Tübingen: Niemeyer.
- Lass, Roger. 1997. *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.
- Maiden, Martin. 2020. Folk Etymology and Contamination in the Romance Languages. *Oxford Research Encyclopaedia in Linguistics*. Oxford: Oxford University Press.
- Masini, Francesca & Jenny Audring. 2019. Construction morphology. In Jenny Audring & Francesca Masini (eds.), *The Oxford Handbook of Morphological Theory*, 365–89. Oxford: Oxford University Press.
- Rainer, Franz. 1993. *Spanische Wortbildungslehre*. Tübingen: Niemeyer.
- Wurzel, Wolfgang U. 1998. On the development of incorporating structures in German. In Richard M. Hogg & Linda van Bergen (eds.), *Historical Linguistics 1995. Volume 2: Germanic Linguistics*, 331–344. Amsterdam & Philadelphia: John Benjamins.





# **ORAL PRESENTATIONS**



---

# Quantitative measures of affix rivalry

*Justine Salvadori*      *Rossella Varvara*      *Richard Huyghe*  
Université de Fribourg    Université de Fribourg    Université de Fribourg

---

## 1 Background

Affix rivalry occurs between affixes that have equivalent semantic functions and can therefore compete in the formation of derivatives (Lindsay & Aronoff, 2013; Arndt-Lappe, 2014; Fradin, 2019; Huyghe & Varvara, 2023; a.o.). However, equivalence may be established only between some of the functions of polyfunctional derivational processes. According to Lieber (2016), for example, the English suffixes *-ation* and *-al* can both derive event (*conversation, portrayal*) and result (*coloration, acquittal*) nouns, but only the former can be used to derive instrument (*decoration*) and agent (*administration*) nouns.

The fact that rival affixes are not always strictly equivalent entails that morphological competition should be considered a gradient relationship. Semantic differences observed between rival affixes can be more or less important, and affixes can be seen as more or less rivaling depending on how close they are semantically. This gradient nature of affix rivalry calls for an appropriate, i.e. quantified, assessment. Ideally, a coefficient of competition should be provided so that different situations of rivalry can be compared both within languages and cross-linguistically.

This work introduces two similarity measures drawn from studies in ecology that can be used to assess degrees of rivalry between polyfunctional affixes: the Sørensen index (Sørensen, 1948), which quantifies how similar two affixes are according to the proportion of functions they share; and the Percentage similarity coefficient (as a complement to the Percentage difference index proposed by Odum, 1950), which quantifies how similar two affixes are considering type frequencies. Two complementary measures — Balanced richness (for the Sørensen index) and Balanced abundance (for the Percentage similarity coefficient) — are also provided to further analyze the semantic dissimilarity between rival affixes. For instance, they can help identify nestedness, i.e. when the functions of an affix A are a subset of the functions of an affix B, and overlap, i.e. when two affixes A and B have functions in common but also specific functions that are not covered by B and A, respectively (Plag, 1999; Guzmán Naranjo & Bonami, 2023; a.o.).

## 2 Case study

French deverbal suffixes often compete for morphosemantic functions (Dubois, 1962; Thiele, 1987; Huyghe & Wauquier, 2021; a.o.). In order to explore the potential of the proposed measures, we selected six of them for a case study: 3 eventive suffixes (*-ade, -ment, -ure*) and 3 agentive suffixes (*-aire, -ant, -eur*). Given that morphological competition can only be investigated through the lexicon, a random sample of 100 French deverbal nouns formed with each suffix was retrieved from the French web corpus FRCOW16A (Schäfer & Bildhauer, 2012; Schäfer, 2015). To identify functions, each collected noun was then semantically analyzed using a double classification (Salvadori & Huyghe, 2023) that distinguishes between the ontological description of the referent (e.g. animate entity, artifact, event) and the relation with the eventuality denoted by the base verb (e.g. agent, instrument, result). In total, 21 ontological and 18 relational classes were considered and assigned to nouns using linguistic tests and

definitions taken from the literature (Flaux & Van de Velde, 2000; Petukhova & Bunt, 2008; Haas et al., 2022; a.o.). The different measures were finally applied to the 6 suffixes based on the 782 word meanings and 37 functions identified in the dataset.

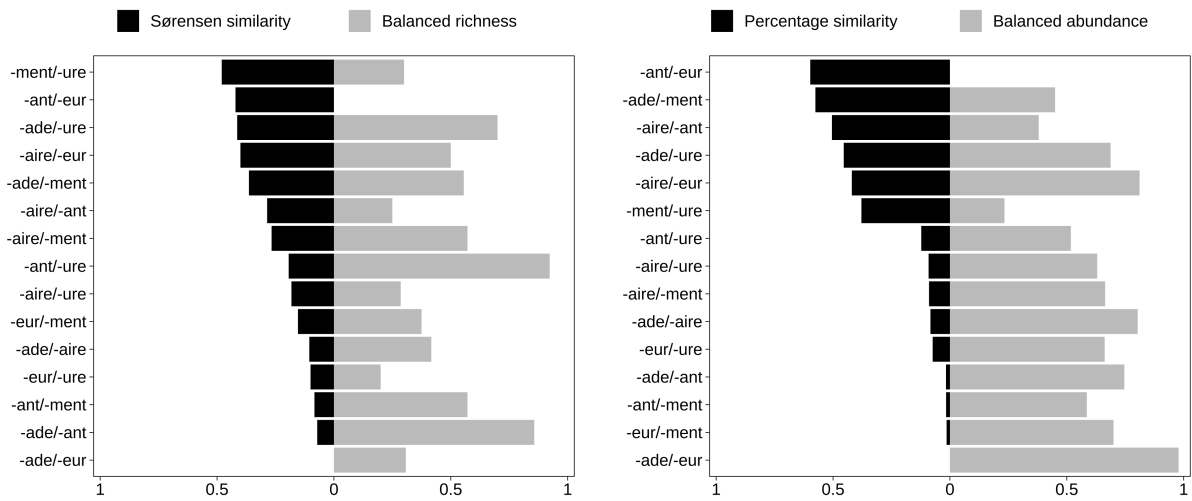


Figure 1: Scores for the incidence- (Sørensen similarity and Balanced richness) and abundance-based (Percentage similarity and Balanced abundance) measures. Pairs of suffixes are ordered from top to bottom by decreasing similarity.

Overall, the results of the case study support the need to approach affix rivalry as a gradient phenomenon. As shown in Figure 1, there are no perfect rivals in the sample and almost all suffixes compete — even in very small proportions. It remains that the pairs composed of suffixes belonging to the same semantic group (i.e. agentive or eventive) obtain higher scores than those contrasting two types of suffixes.

The proposed measures highlight different facets of similarity relationships and complement each other accordingly. As incidence-based measures, the Sørensen and Balanced richness indices allow in-

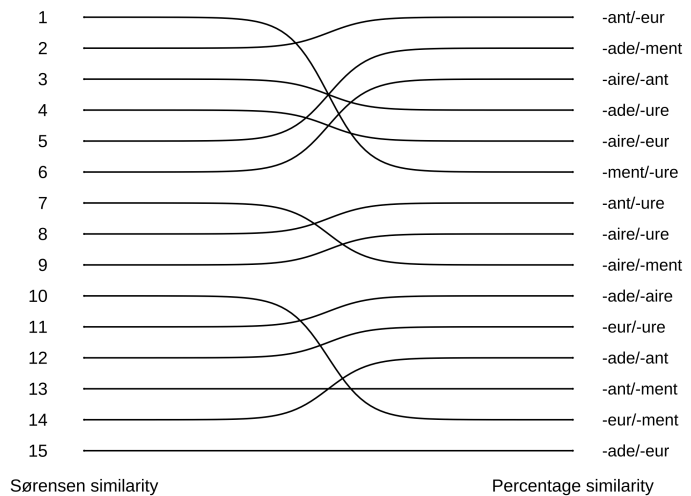


Figure 2: Ranking of the suffix pairs according to the Sørensen vs. Percentage similarity measures.

depth investigation of functionality structures. As abundance-based measures, the Percentage similarity and Balanced abundance indices can weight functional rivalry by realization frequency and shed a different light on the sharing of functions. The comparison between the two types of measures informs on the architecture of rivalries and on the (in)congruence between the number of shared functions and the number of derivatives that instantiate these functions. In this sample, there is a strong and significant correlation between the Sørensen and the Percentage similarity scores (Mantel test:  $r = .868$ ,  $p < .01$ ). This suggests that suffixes that have many functions in common also tend to present a relatively even distribution of derivatives across shared functions, although some exceptions can be noted. For example, while the suf-

fixes *-ment* and *-ure* are the most similar according to the Sørensen index, they lose 5 places in the ranking based on the Percentage similarity measure (see Figure 2), meaning that, although they share a high number of functions, they realize them at different frequencies.

### 3 Conclusion

This work introduces different 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 presented in the study should be considered a first step toward a comprehensive measurement of morphological competition. They do not account for the diachronic evolution and change in productivity that can affect rivalry in the long run, nor do they inform about the availability of an affix when coining new words at a given point in time. In the future, these similarity indices could be examined diachronically and could also be combined with productivity measures to improve the assessment of rivalry.

### References

- Arndt-Lappe, Sabine. 2014. Analogy in suffix rivalry: the case of English *-ity* and *-ness*. *English Language & Linguistics* 18(3). 497–548.
- Dubois, Jean. 1962. *Étude sur la dérivation suffixale en français moderne et contemporain : essais d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Larousse.
- Flaux, Nelly & Danièle Van de Velde. 2000. *Les noms en français : esquisse de classement*. Editions Ophrys.
- Fradin, Bernard. 2019. Competition in derivation: what can we learn from French doublets in *-age* and *-ment*? In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds.), *Competition in inflection and word-formation*, 67–93. Berlin: Springer.
- Guzmán Naranjo, Matías & Olivier Bonami. 2023. A distributional assessment of rivalry in word formation. *Word Structure* 16(1). 86–113.
- Haas, Pauline, Lucie Barque, Richard Huyghe & Delphine Tribout. 2022. Pour une classification sémantique des noms en français appuyée sur des tests linguistiques. *Journal of French Language Studies* 1–30. doi:<http://doi.org/10.1017/S0959269522000187>.
- Huyghe, Richard & Rossella Varvara. 2023. Affix rivalry: theoretical and methodological challenges. *Word Structure* 16(1). 1–23. doi:<https://doi.org/10.3366/word.2023.0218>.
- Huyghe, Richard & Marine Wauquier. 2021. Distributional semantics insights on agentive suffix rivalry in French. *Word Structure* 14(3). 354–391.
- Lieber, Rochelle. 2016. *English nouns: the ecology of nominalization*, vol. 150. Cambridge University Press.
- Lindsay, Mark & Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. *Morphology in Toulouse: Selected Proceedings of Décembrettes* 7. 133–153.
- Odum, Eugene P. 1950. Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology* 31(4). 587–605.
- Petukhova, Volha & Harry Bunt. 2008. LIRICS semantic role annotation: design and evaluation of a set of data categories. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 39–45. European Language Resources Association.

- Plag, Ingo. 1999. *Morphological productivity: structural constraints in English derivation*. Berlin: Walter de Gruyter.
- Salvadori, Justine & Richard Huyghe. 2023. Affix polyfunctionality in french deverbal nominalizations. *Morphology* 33(1). 1–39.
- Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lungen & Andreas Witt (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, 28–34. Institut für Deutsche Sprache.
- Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the Web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. European Language Resources Association.
- Sørensen, Thorvald 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. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter* 5. 1–34.
- Thiele, Johannes. 1987. *La formation des mots en français moderne*. Presses de l'Université de Montréal.

---

## A story-based approach to derivational paradigms

---

Daniele Sanacore

Université Toulouse - Jean Jaurès

Nabil Hathout

Université Toulouse - Jean Jaurès

Fiammetta Namer

Université de Lorraine

---

**Derivational family slicing.** The paradigmatic nature of derivation has been much discussed in the last decades, in the perspective of unifying inflection and derivation (Van Marle, 1985; Stump, 1991; Bauer, 1997; Boyé & Schalchli, 2016; Hathout & Namer, 2019). To this end, many authors have proposed to extend paradigms to derivation (Bochner, 1993; Booij, 2010; Jackendoff & Audring, 2018; Hathout & Namer, 2022). Despite the growing number of papers adopting a paradigmatic approach to derivation, the nature of derivational paradigms is still debated. Bonami & Strnadová (2019) propose that derivational paradigms are alignments of “slices” of derivational families (that we may call “paradigmatic families”) having the same content relations. Like Bauer (2019) and Antoniova & Štekauer (2016), they consider that the structure of paradigms is determined by meaning. On the other hand, the question of the delimitation of the derivational paradigms has hardly been discussed. In this talk, we focus on this question. We propose a methodology for the slicing of derivational families into paradigmatic families that can be aligned in order to form derivational paradigms. In this abstract, we illustrate this methodology with French examples.

**Stories that tell morphosemantic relations.** Our procedure starts from a derivational family. As an example, consider the French family of the artifact noun *pot* ‘pot’ in (1).

- (1)  $F1 = \{pot, poterie, potier, rempoter, repotage\}$   
‘pot’, ‘pottery’, ‘potter’, ‘to repot’, ‘repotting’

In order to identify all the relevant semantic relations in  $F1$ , we first consider all  $F1$  subsets of size  $\geq 2$ . We refer to this cover as  $cov(F1)$  as in (2).

- (2)  $cov(F1) = \{\{pot, poterie\}, \{pot, potier\}, \dots, \{rempoter, repotage\}, \{pot, potier, poterie\}, \dots, \{pot, poterie, potier, rempoter, repotage\}\}$

A first difficulty when it comes to identifying the semantic relations that connect the lexemes in a subset or are involved in the characterization of these relations is the lack of resources which a systematic description of the lexical relations present in the lexicon could be drawn from. Note that resources such as WordNet (Fellbaum, 1999), FrameNet (Ruppenhofer et al., 2003) or even JeuxDeMots (Lafourcade & Joubert, 2013) would not be suitable because the range of the relations they provide is too limited. Another option would be to interview speakers to obtain such descriptions, by asking them to tell us a story that contains the words in the subset (as in some radio games). Unfortunately, we do not have the means to carry out such large-scale surveys. For this reason, our proposal is purely methodological<sup>1</sup>. In order to illustrate our method and unfold its different stages, we propose to implement it on some stories that we will produce ourselves.

For each subset in  $cov(F1)$ , we produce a set of stories that contain instances of the lexemes included in the subset, like the ones in (3) for the subset  $\{pot, poterie\}$ . Stories may be made up of one sentence (3a) or many ones (3b).

- (3) a. h11 = Hier, Marc a fabriqué un pot magnifique dans le cours de poterie.<sup>2</sup>

---

<sup>1</sup>The availability of generative models like *ChatGPT* makes it possible to envisage a large-scale production of the stories we need.

<sup>2</sup>‘Yesterday, Marc made a beautiful pot in the pottery class.’

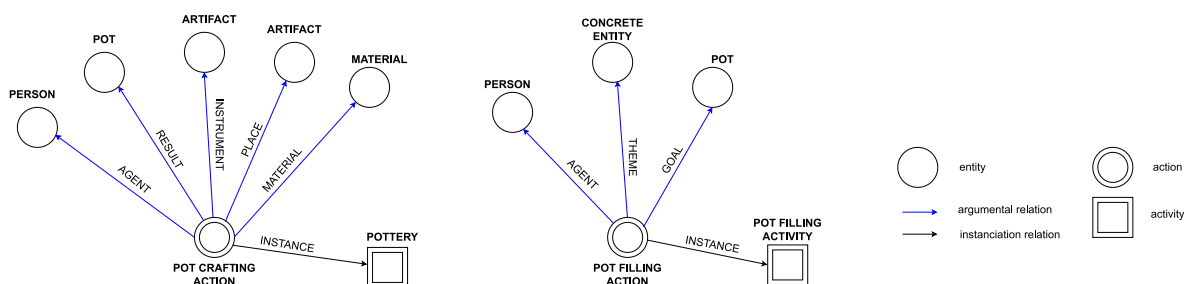


Figure 1: Family meaning bundles (FMB) built from stories about the lexemes in word family F1. The FMB on the left hand side describes the action of crafting pots. The one on the right hand side describes the action of transferring to a new pot. The entities are represented by their ontological categories. In the FMB on the left hand side, the action of making pots is an instance of the activity pottery. In the bundle on the right, the action of transferring into a new pot is an instance of an activity of the same nature, which may be usually performed in gardening.

- b. h12 = Ambre a toujours voulu apprendre à fabriquer des pots. Le mois dernier elle s'est inscrite à un cours de poterie proposé par la mairie.<sup>3</sup>

Our hypothesis is that lexemes that are strongly linked by semantic relations will regularly co-occur in stories. For instance, the lexemes *pot*, *potier* and *poterie* are strongly semantically related and will regularly co-occur in stories. On the other hand, we expect that lexemes that are semantically distant may co-occur in fewer stories and that their relations in these stories will be quite episodic. For example, a story where someone takes a plant out of a pot made up for the subset {*pot*, *rempoter*}, like the one in (4) will hardly involve a potter or pottery (the activity of pot crafting).

- (4) h13 = Nous avons dû rempoter notre Aloe Vera parce que son ancien pot est devenu trop petit.<sup>4</sup>

**Abstracting semantic bundles from the stories.** To turn stories like h11, h12 or h13 told in a textual form into more formal objects that can be more easily manipulated and compared, we propose to transcribe them as semantic networks similar to those proposed by Sowa (2014). The operation is performed separately for each subset of  $cov(F1)$ . Then, we replace the vertices of the network by labels that indicate the ontological class of the corresponding referents. The graphs are then linearized and the relations they contain are clustered. The resulting clusters are graphs that describe the semantic content that can be associated to the lexemes of the subset. We will call this graph “subset meaning bundle.” (SMB). The operation is repeated for all subsets of  $cov(F1)$  to obtain a set of subset meaning bundles that we then align on the basis of the ontological categories of the entities they contain. The more general meaning bundles obtained in this way may be called “family meaning bundles” (FMB). These can in turn be aligned in the same way as before to build semantic paradigms similar to those proposed by Hathout & Namer (2022). We will call them “lexical meaning bundles.” (LMB) Figure 1 shows two family meaning bundles that describe actions and activities (Roché, 2017; Fradin, 2020) that involve pots: one where a pot is the result of the action and the other where it is a goal.

Note that stories about subsets of lexemes from a family like F1 do not necessarily mention all the concepts contained in one of the FMB. For example, the story in (3a) does not speak of

<sup>3</sup>Ambre has always wanted to learn how to make pots. Last month she signed up for a pottery course offered by the city hall.’

<sup>4</sup>‘We had to repot our Aloe Vera because its old pot became too small.’



instruments or materials. Conversely, some concepts included in the FMB may have no realization in the word families they originate from. For example, the material in the FMB on the left hand side in Figure 1 is not realized by a lexeme included in F1.

**Family meaning bundle alignment.** The method illustrated on the family of *pot* can be applied to the other derivational families. Each family yields one or several family meaning bundles that may be aligned with FMB from other families. The alignment is based on the ontological nature of their entities and events. For example the family of *brique* (5) includes lexemes denoting artifacts (*brique*), people that make these artifacts (*briquetier*) and places where the artifacts are made (*briqueterie*). It yields a FMB that may be aligned with the FMB on the left hand side in Figure 1 which has these same vertices. Similarly, the family of *bouteille* in (6) yields a FMB describing a bottle filling (*embouteiller*; *embouteillage*) that may be aligned with the one on the right hand side in Figure 1. In this way, lexical meaning bundles are semantic paradigms that delimit and structure the derivational paradigms.

(5)  $F2 = \{brique, briquetier, briqueterie\}$   
 ‘brick’, ‘brickmaker’, ‘brick factory’

(6)  $F3 = \{bouteille, embouteiller, embouteillage\}$   
 ‘bottle’, ‘to bottle’, ‘bottling’

**Meaning bundle projection.** We now can slice word families into paradigmatic families by projecting on them the lexical meaning bundles. For example, the projection of LMB having the same structure as the FMB in Figure 1 on the derivational family F1 results in the paradigmatic families in (7). We see that two paradigmatic families overlap and share the lexeme *pot*. Paradigmatic families in (7a) and (7b) highlight two facets of the meaning of *pot*: its production and its use.

(7) a.  $f_1 = (pot, potier, poterie)$   
 b.  $f_2 = (pot, rempoter, rempotage)$

Paradigmatic families align in semantically delimited derivational paradigms. Table 1 presents the derivational paradigm related to artifact making, and the Table 2 the one related to moving entities into recipients. The ontological and relational labels of the semantic bundle serve as indexes of the paradigm columns. We can see in Table 1 that some concepts (vertices) in the meaning bundles may not be realized morphologically in some families. It is the case for the materials which are not morphologically realized in the families of *pot* and *brique* but is in the family of *fer-blanc* ‘tinplate’, *ferblanterie* ‘tinware’, *ferblantier* ‘tinsmith’. We also see that one concept in a lexical meaning bundle may correspond to more than one lexeme in a family. Both the verb *rempoter* and the action noun *rempotage* correspond to the event node *pot* filling action in the right FMB in Figure 1. Similarly, several lexemes in a paradigmatic family may have the same form, as in the case of *poterie* and *ferblanterie* (activity and artifact).

artifact	person	activity	place	material
<i>pot/poterie</i>	<i>potier</i>	<i>poterie</i>	-	-
<i>brique</i>	<i>briquetier</i>	-	<i>briqueterie</i>	-
<i>ferblanterie</i>	<i>ferblantier</i>	<i>ferblanterie</i>	-	<i>fer-blanc</i>
...	...	...	...	...

Table 1: Derivational paradigm of artifact making families

recipient	filling <sub>V</sub>	filling <sub>N</sub>
<i>pot</i>	<i>rempoter</i>	<i>rempotage</i>
<i>bouteille</i>	<i>embouteiller</i>	<i>embouteillage</i>
...	...	...

Table 2: Derivational paradigm of recipient filling families

## References

- Antoniova, Vesna & Pavol Štekauer. 2016. Derivational paradigms within selected conceptual fields—contrastive research. *Facta Universitatis, Series: Linguistics and Literature* 61–75.
- Bauer, Laurie. 1997. Derivational paradigms. In *Yearbook of morphology 1996*, 243–256. Springer.
- Bauer, Laurie. 2019. Notions of paradigm and their value in word-formation. *Word Structure* 12(2). 153–175.
- Bochner, Harry. 1993. *Simplicity in generative morphology*. De Gruyter Mouton.
- Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2). 167–197.
- Booij, Geert. 2010. Construction morphology. *Language and linguistics compass* 4(7). 543–555.
- Boyé, Gilles & Gauvain Schalchli. 2016. The status of paradigms. *The Cambridge handbook of morphology* 206–234.
- Fellbaum, Christiane (ed.). 1999. *Wordnet: an electronic lexical database*. Cambridge, MA: MIT Press.
- Fradin, Bernard. 2020. Characterizing derivational paradigms. In *Paradigmatic relations in word formation*, 49–84. Brill.
- Hathout, Nabil & Fiammetta Namer. 2019. Paradigms in word formation: what are we up to? *Morphology* 29(2). 153–165.
- Hathout, Nabil & Fiammetta Namer. 2022. Paradis: a family and paradigm model. *Morphology* 1–43.
- Jackendoff, Ray & Jenny Audring. 2018. Relational Morphology in the Parallel Architecture. In *The Oxford Handbook of Morphological Theory*, 390–408. Oxford University Press.
- Lafourcade, Mathieu & Alain Joubert. 2013. Bénéfices et limites de l’acquisition lexicale dans l’expérience jeuxdemots. In Nuria Gala & Michael Zock (eds.), *Ressources lexicales: Contenu, construction, utilisation, évaluation*, vol. 30 *Linguisticae Investigationes, Supplementa*, 187–216. Amsterdam: John Benjamins.
- Roché, M. 2017. Les familles dérivationnelles: comment ça marche. *Toulouse: Université Toulouse 2*.
- Ruppenhofer, Josef, Collin F. Baker & Charles J. Fillmore. 2003. The framenet database and software tools. In *Proceedings of the tenth euralex international congress*, 371–375. Copenhagen, Denmark.
- Sowa, John F. 2014. *Principles of semantic networks: Explorations in the representation of knowledge*. Morgan Kaufmann.
- Stump, Gregory T. 1991. A paradigm-based theory of morphosemantic mismatches. *Language* 675–725.
- Van Marle, Jaap. 1985. *On the paradigmatic dimension of morphological creativity*. De Gruyter.

---

# High frequency derived words have low semantic transparency mostly only if they are polysemous

*Martha Booker Johnson*

The Ohio State University

*Micha Elsner*

The Ohio State University

*Andrea D. Sims*

The Ohio State University

---

The development of word vectors as an implementation of distributional semantics (Boleda, 2020, *inter alia*) offers new tools for quantitatively testing old ideas about morphology. Since Bybee (1985), an often-repeated claim is that a strong relationship holds between a derived lexeme’s token frequency and its semantic relationship to its base. Specifically, high frequency is posited to facilitate low semantic transparency as a function of lexical storage (Baayen, 1993; Bybee, 1985). Yet surprisingly little work has tested this using a quantitative measure of semantic transparency. Closest is Hay (2001), who treats semantic transparency as binary. We start from the observation that polysemy complicates base-derivative relations (e.g. Lapesa et al., 2018; Salvadori & Huyghe, 2023). An open question thus has to do with the role that polysemy plays in the relationship between derivative frequency and semantic transparency, if any. We use word vectors to test for a correlation between semantic transparency and derivative frequency in English, examining the role of polysemy.

We present three analyses. First, using a large dataset we show that the simple claim of an inverse relationship between derivative frequency and semantic transparency (operationalized as cosine similarity) is not supported, contrary to received wisdom. Second, using a subset of the data we show that the expected relationship *can* be detected, but only when interactions between frequency and polysemy are considered. Finally, we validate this result by showing that similar polysemy effects also emerge in human judgments of the semantic relatedness of bases and derivatives. Specifically, high polysemy derivatives exhibit an inverse relationship between derivative frequency and semantic transparency but low polysemy derivatives do not.

In short, polysemy mediates the relationship between frequency and semantic transparency, a fact that has not been sufficiently recognized in previous work.

## 1. No simple correlation between frequency and semantic transparency

To test for a correlation between derivative frequency and semantic transparency in the English lexicon, we started with 10,465 derived English lexemes from Sims & Parker (2015), which correspond to all of the lexemes in CELEX (Baayen et al., 1995) that end in one of 54 English derivational suffixes. We extracted these lexemes’ bases from CELEX’s morphological analysis. Lemma frequency for each derivative and base word was calculated from the training set of the Tensorflow Wiki40b dataset (Guo et al., 2020), which provides English Wikipedia data cleaned of extraneous text. We lemmatized and part-of-speech (POS) tagged the dataset using CoreNLP (Manning et al., 2014) and then calculated token frequency for each lemma. For all models we converted frequency counts to log instances per million words of corpus (log ipm) since word frequencies are Zipfian. Base-derivative pairs in which either lemma had fewer than 8 tokens (= 0.1 ipm) were removed because word vectors are typically unstable for low-frequency items. Suffixes with fewer than 10 example pairs were then also dropped. This resulted in a dataset containing 3,286 base-derivative pairs for 34 suffixes.

We operationalize semantic transparency as the cosine similarity of a derived lexeme’s vector to its base lexeme’s vector. For each base and derived lexeme, we retrieved its 300-dimensional vector from Fares et al.’s (2017) lemmatized English model that was trained on

the English Wikipedia dump of February 2017. We then calculated cosine similarity (ranging between 0 and 1, with higher values indicating greater semantic transparency) for each base-derivative pair. (Cosine similarity was chosen as a measure in order to maximize comparability to the task in Analysis 3, which asked participants to compare the similarity of base and derivative forms.)

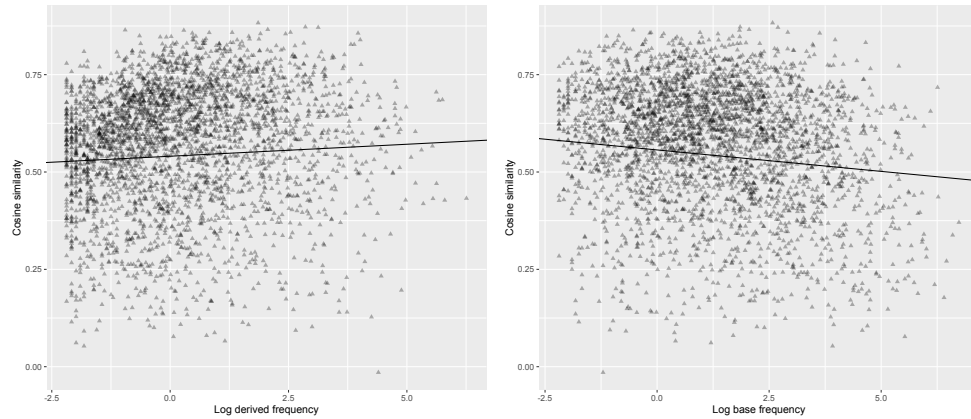


Figure 1: (Lack of) correlation between derived frequency (in log ipm) and cosine similarity (left panel), or base frequency (in log ipm) and cosine similarity (right panel)

As shown in Figure 1, the data are widely dispersed. We constructed a mixed effects regression model that predicted cosine similarity from derived frequency, with suffix and base as random intercepts. A negative relationship between derivative frequency and cosine similarity was expected. Results show frequency was significantly *positively* correlated with cosine similarity ( $\beta = 0.006$ ,  $t(3277) = 3.766$ ,  $p < 0.001$ ) but had an extremely low marginal  $R^2 = 0.0017$ , indicating that word frequency accounts for almost none of the variance. Thus, no simple, robust correlation between low semantic transparency and high frequency is observed. (A model with base frequency as a predictor produced a similarly weak pattern in the opposite direction.)

## 2. A correlation exists, but predominantly for highly polysemous derived words

Analysis 1 suggested that the null result was caused by uncontrolled variables, with polysemy as the main suspect. To investigate this we used 109 base-derivative pairs (a subset of the Analysis 1 data) that had been experimental stimuli for a ratings task (McKenzie, 2019). Derivative frequency and base frequency were strongly correlated, so to use both as predictors in the model we residualized base frequency on derived frequency. The number of senses of the derivative and of the base (our measures of polysemy) was calculated as the number of senses listed in the online Oxford English Dictionary (oed.com). A final, stepped-down mixed effects regression model had the following fixed effects: derived frequency, squared derived frequency, residualized base frequency, squared residualized base frequency, and derived number of senses. There was one two-way interaction: residualized base frequency\*derived number of senses. There was a random intercept for affix. All factors were centered or sum-contrasted, as appropriate.

The left panel of Figure 2 visualizes the main effect (and quadratic) relationship between derivative frequency and cosine similarity, in the expected direction (i.e. a negative correlation). Even more interesting is the two-way interaction shown in the right panel. Since base frequency was residualized on derived frequency, an x-axis value of 0 represents a base that is exactly as frequent as would be expected given the frequency of the derived word. Negative values indicate base-derivative pairs in which base frequency is lower than expected (or equivalently, derived frequency is higher). The values for derived number of senses are the mean

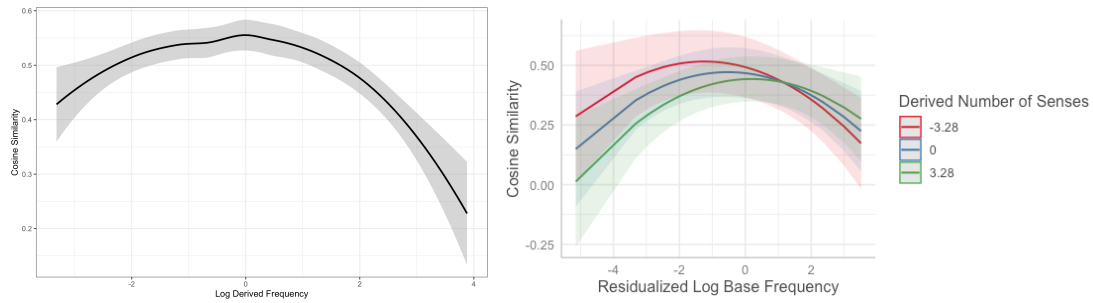


Figure 2: Model-predicted values for cosine similarity based on derived frequency (left panel) and interaction between residualized base frequency and derived number of senses (right panel)

(blue) and one standard deviation above (green) and below (red) the mean. As shown, among derivatives with higher-than-expected frequency, lower cosine similarity values are observed for those that are also highly polysemous. The correlation between low semantic transparency and high frequency is thus more strongly a property of highly polysemous derived words.

### 3. A similar pattern emerges in similarity judgments

Finally, as a check on whether the word vectors are adequately capturing human intuitions about semantic relatedness of base-derivative pairs, we reanalyzed data from McKenzie (2019), which asked 24 native speakers of English to provide semantic similarity judgments for the same 109 pairs used for Analysis 2. Participants responded to the prompt “How similar is the meaning of the word [DERIVED] to the meaning of the word [BASE]?” using a continuous scale. The fixed effects in the final, stepped-down model were derived frequency, residualized base frequency, derived number of senses, and base number of senses. The final model also included three two-way interactions — derived frequency\*derived number of senses, residualized base frequency\*derived number of senses, and derived number of senses\*base number of senses — and random intercepts for participant and word, with affix as a grouping factor for word. All factors were centered or sum-contrasted, as appropriate.

The relationship between derived frequency, derived number of senses and base number of senses is visualized in Figure 3. For words with few derived senses (red line), there is no change in response based on derived frequency. For derived words with average and above average number of senses (blue and green lines), however, participant similarity judgments decrease as derived frequency increases, with a steeper slope for words with more senses. Thus, similarly to what was observed with cosine similarity, a negative relationship between derivative frequency and semantic transparency is characteristic only of polysemous derivatives. Additionally, an interaction between base polysemy and derivative polysemy is observed. For low polysemy derivatives (red line), as the polysemy of the base increases, semantic transparency judgments decrease. Thus, polysemy of both the derivative and the base affects judgments.

Vector models of word-formation have proliferated recently, showing that the distributional semantic approach can be profitably applied to a range of morphological questions. Our study contributes to this line of research. Specifically, we show that frequent claims suggesting that high word token frequency is straightforwardly correlated with low semantic transparency do not hold in English. Instead, the relationship is crucially mediated by polysemy. The most semantically opaque derivatives have high frequency **and** are highly polysemous. Thus, despite being received wisdom, the relationship between frequency and semantic transparency (in English) is more complex than previously understood. Ongoing work includes implementation

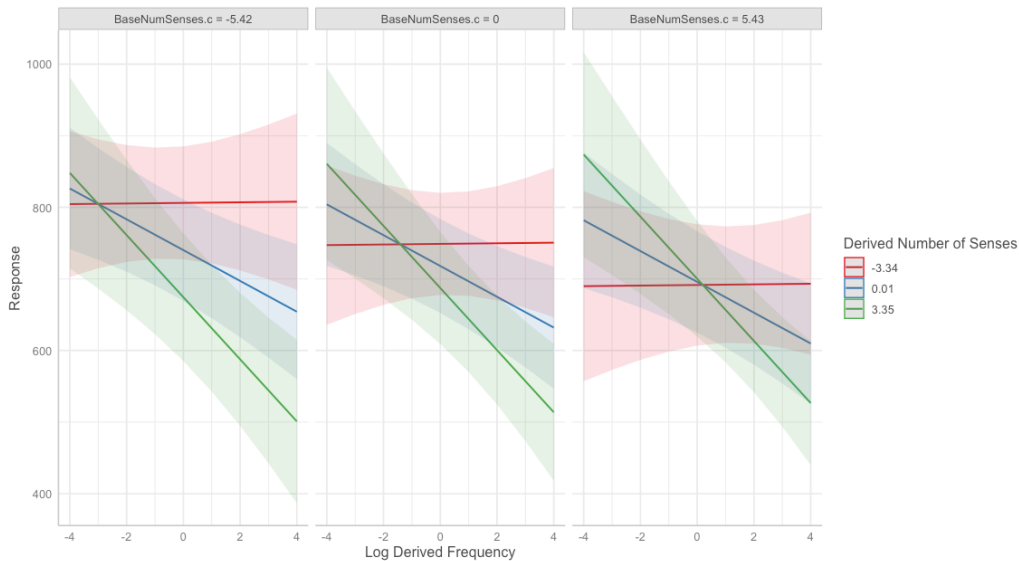


Figure 3: Model-predicted values for similarity judgments, with derived frequency (x-axis), derived number of senses (series; red = low, green = high), and base number of senses (panels)

of Marelli & Baroni’s (2015) measure of semantic transparency, which calculates compositional vectors, to take account of the affix’s semantic contribution, as well as expanding the dataset (Analysis 2) to include polysemy information for a larger number of base-derivative pairs.

## References

- Baayen, R. H. 1993. On frequency, transparency, and productivity. In G. Booij & J. van Marle (eds.), *Yearbook of morphology 1992*, 181–208. Kluwer.
- Baayen, R. H., R. Piepenbrock & L. Gulikers. 1995. The CELEX Lexical Database (CD-ROM).
- Boleda, G. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6. 213–234.
- Bybee, J. 1985. *Morphology: A study of the relation between meaning and form*. John Benjamins.
- Fares, M., A. Kutuzov, S. Oepen & E. Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, 271–276.
- Guo, M., Z. Dai, D. Vrandečić & R. Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. *Proceedings of LREC 2020: 12th International Conference on Language Resources and Evaluation* 2440–2452.
- Hay, J. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39. 1041–1070.
- Lapesa, G., L. Kawaletz, I. Plag, M. Andreou, M. Kisselew & S. Padó. 2018. Disambiguation of newly derived nominalizations in context: A Distributional Semantics approach. *Word Structure* 11. 277–312.
- Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard & D. McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 55–60.
- Marelli, M. & M. Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122. 485–515.
- McKenzie, M. 2019. Effects of relative frequency on morphological processing in Russian and English. BA thesis. The Ohio State University.
- Salvadori, J. & R. Huyghe. 2023. Affix polyfunctionality in French deverbal nominalizations. *Morphology* 33. 1–39.
- Sims, A. & J. Parker. 2015. Lexical processing and affix ordering: Cross-linguistic predictions. *Morphology* 25. 143–182.

---

# Paralex: a DeAR standard for rich lexicons of inflected forms.

Sacha Beniamine<sup>1</sup>, Cormac Anderson<sup>2</sup>, Mae Carroll<sup>3</sup>, Matías Guzmán Naranjo<sup>4</sup>,  
Borja Herce<sup>5</sup>, Matteo Pellegrini<sup>6</sup>, Erich Round<sup>1</sup>, Helen Sims-Williams<sup>1</sup>, Tiago Tresoldi<sup>7</sup>

<sup>1</sup>University of Surrey; <sup>2</sup>Max Planck Institute EVA; <sup>3</sup>Australian National University;

<sup>4</sup>Albert-Ludwigs-Universität Freiburg; <sup>5</sup>University of Zurich;

<sup>6</sup>CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan; <sup>7</sup>Uppsala Universitet

---

## 1 Introduction

We present Paralex<sup>1</sup>, a new technical standard for inflected lexicons in tabular format. Inflected lexicons document the inflected forms of words, such as the conjugations of verbs and the declensions of nouns. Such datasets are crucial to support morphological investigation using both computational and traditional methodologies, and constitute a necessary foundation for data-driven studies of individual systems as well as large scale typological works.

Many existing morphological datasets are not published durably. Some use proprietary formats, others exist solely as web portals (see Maiden et al. 2010, standardised by Beniamine et al. 2019). Resources which were intended for manual exploration are often not machine-readable. Those which are, often use their own sets of conventions, and are not inter-operable (see eg. Bonami et al., 2014; Pellegrini & Passarotti, 2018; Feist & Palancar, 2015). The Unimorph datasets (McCarthy et al., 2020) do provide inter-operable lists of inflected forms, but their usefulness for linguistic investigation is limited both by their automatic extraction and an exclusive focus on orthography (Malouf et al., 2020).

The Paralex standard aims to bring about high quality resources, which can be richly annotated, are machine-readable, inter-operable and durable. It describes lexicons constituted of csv tables in long format forming a relational database, accompanied by metadata in json format (§ 2). The standard is devised to promote good data practices and abides by the FAIR principles (Wilkinson et al., 2016), as well as our own set of principles (DeAR, § 3).

## 2 Data and metadata formats

Paradigms are conventionally written as tables in a variety of formats (Corbett, 2013). Authors often present single paradigms as in Table 1.a., where rows and columns represent morpho-syntactic features. Such tables are impractical for presenting many lexemes, as this would require multiple tables. Thus, a more common format for this purpose (see e.g. Flexique, Bonami et al. 2014) is the Plat (Stump & Finkel, 2013), which arranges paradigm cells in columns, and lexemes in rows, as in Table 1.b. This format is more generally known as a *wide form* table. The major draw-back of this format is that it can only ever express a single piece of information per cell/lexeme intersection, making it impossible to cleanly record overabundant forms (Thornton, 2012) or multiple pieces of information for each form, including but not limited to its phonological form, its frequency, source, analysis, etc. Thus, we adopt instead the *long form* (For more discussion on wide vs long form for linguistic data, see Forkel et al., 2018), in which each inflected wordform is given its own row, as shown in Table 1.c. Rows have unique identifiers, and columns for forms, cells, and lexemes, and any further information. Overabundant word forms lead to multiple rows.

---

<sup>1</sup>The full standard specifications and documentation can be found at <https://www.paralex-standard.org>

(a) Single paradigm table

	SINGULAR	PLURAL
NOMINATIVE	rosa	rosae
VOCATIVE	rosa	rosae
ACCUSATIVE	rosam	rosās
GENITIVE	rosae	rosārum
DATIVE	rosae	rosīs
ABLATIVE	rosā	rosīs

(c) Long form table

form_id	cell	lexeme	orth_form
f1	NOM.SG	rosa	rosa
f2	VOC.SG	rosa	rosa
f3	ACC.SG	rosa	rosam
f13	NOM.SG	dominus	dominus
f14	VOC.SG	dominus	domine
...	...	...	...

(b) Wide form table

lemma	NOM.SG	VOC.SG	ACC.SG	GEN.SG	DAT.SG	ABL.SG	NOM.PL	voc.pl	...
ROSA	rosa	rosa	rosam	rosae	rosae	rosā	rosae	rosās	...
DOMINUS	dominus	domine	dominum	dominī	dominō	dominō	dominī	dominī	...

Table 1: Paradigm formats, illustrated on two Latin nouns (Pellegrini &amp; Passarotti, 2018).

A paralex lexicon is minimally constituted of a simple forms table (see Table 1.c), associating forms (orthographic or phonological) with paradigm cells, lexemes, and unique identifiers. The standard further describes tables to document entities from the forms table: lexemes, cells, feature-values, sounds, and graphemes. A tags table declares user-defined properties of forms and a very flexible frequencies table records frequency measurements. A set of columns is pre-defined for each table. Paralex lexicons may use pre-defined tables and columns, adding any additional ones as needed. The tables are linked by two types of relationships. Foreign key relations allow direct references between tables rows: For example, “NOM.SG” in the cell column of the forms table refers to the row with the identifier “NOM.SG” in the cells table. The foreign key relations between the three main tables are illustrated in Figure 1. Moreover, elements from some tables are composed of identifiers from other tables. For example, cells (e.g. NOM.SG) are composed of feature-values separated by dots (NOM, SG), orthographic forms are composed of graphemes, phonological forms are composed of sounds symbols, etc.

Beyond relations between tables, references to linked vocabularies greatly increase the value of datasets, and are encouraged. Languages can be denoted by glottocodes or ISO-639-2 codes; cells and features can refer to the Universal Dependencies and/or to the Unimorph conventions, sounds may refer to CLTS’ BIPA (Anderson et al., 2018), etc. To further enhance interoperability with different resources, the standard is coupled with an ontology, where RDF classes and properties are introduced, corresponding to tables and columns defined in the standard, respectively. Their relation to existing standard vocabularies – such as the General Ontology for Linguistic Description (GOLD; Farrar & Langendoen 2003) and the Lexicon Model for Ontologies (OntoLex; McCrae et al. 2017) – is expressed by means of sub-class (`subclassOf`) and sub-property (`subPropertyOf`) relations, as defined in the RDF Schema vocabulary. This allows the conversion of Paralex data into ontolox-compliant lexicons in RDF, guaranteeing semantically richer interoperability not only with other morphological lexicons, but also with lexical resources of other kinds.

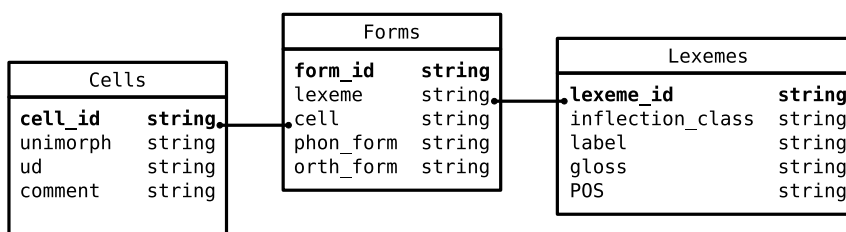


Figure 1: Relations between the three main Paralex tables.



Metadata are any information about the dataset that are not directly part of the data. A first type of metadata is global information about a dataset, such as its author(s), name, identifier, license, etc. This information is usually provided in landing pages, articles or documentation files, in a way that is easy to understand for humans, but often not machine readable. Furthermore, many other pieces of information about the data itself are often left implicit, such as: what does each table document? How are tables related? What values are expected in each column? This is neither future-proof (the context is likely to be lost) nor machine-readable. Thus, Paralex lexicons explicitly encode metadata in a `json` file following the frictionless standard (Fowler et al., 2018). Its creation is facilitated by a Paralex python package which can fill in all conventional information from the standard.

### 3 Philosophy

Paralex datasets adhere to the FAIR principles (Wilkinson et al., 2016), which focus on data users: they ensure that datasets be readable by both machines and by humans across sub-fields, disciplines and time. Focusing on data creators, we introduce our own set of principles for scientific data: **D**ecentralisation, **A**utomated verification and **R**evisable pipelines (DeAR).

The acronym FAIR stands for Findable, Accessible, Interoperable and Reusable. Findability relies on persistent global identifiers (F1), rich metadata (F2) referring to the identifier (F3), and indexation in searchable resources (F4). Paralex addresses F2 through the metadata file, recommends using DOIs (F1,F3), and archiving lexicons in dedicated repositories (F4). These measures also ensure that the data is Accessible. Inter-operability consists in using a formal, accessible, shared, broadly applicable language for knowledge representation (I1), FAIR vocabularies (I2) and reference to other (meta)data (I3). The formats chosen for Paralex fit the descriptions in I1. Compliance with I2 and I3 rely on the use of linked identifiers, and the Paralex ontology. Finally, rich metadata also addresses reusability, by ensuring well-described data (R1) which can be re-used and combined in other contexts.

When faced with the task of creating a large number of standardised datasets, one solution is for a single team to retro-standardise large amounts of data into a single database. Unfortunately, compounded datasets tend to be cited at the expense of original resources, leading to loss of recognition for data creators. Moreover, centralisation concentrates power over indigenous and endangered languages into the hands of a few institutions, going contrary to the CARE principles (Carroll et al., 2020). Thus, Paralex rather aims to stimulate a **D**ecentralized adoption of the standard. Although there must of course be a single definition of the standard, we intend to make it easy and flexible to use, and to produce tools which incentivize its adoption. Creating large databases is difficult and error-prone. In order to improve data quality, we promote the **A**utomated validation of datasets. The statements contained in the metadata file can be verified automatically against the data using existing frictionless tools. This process can ensure perfect formatting, valid references across tables, and check expected properties of data content. Validation can be performed at each update of the data to maintain high data quality. Finally, it is crucial for data to be linked to its published presentations (such as websites) through **R**evisable pipelines. The inter-operability of standardized datasets makes it possible to create websites which can be re-generated whenever the data is updated.

### 4 Conclusion

The Paralex standard provides formal conventions for coding inflected lexicons and their metadata. It is suited to encoding inflectional systems across languages, for purposes ranging from

lists of inflected forms to richly annotated lexicons. It describes mechanisms to handle phenomena such as overabundance, defectivity, and multiple types of variation. It is accompanied by a helper tool to generate the metadata with minimal effort, and a tool is in development to create static websites automatically. As linguists, we are most interested in the parts of language that are complex to analyze, and thus complex to code. Thus, this standard accommodates a great deal of flexibility regarding the exact content of the data, allowing linguists to make project-specific analytical choices about content, while reaping other benefits of standardization.

## References

- Anderson, Cormac, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel & Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting* 4(1). 21–53. doi:10.2478/yplm-2018-0002.
- Beniamine, Sacha, Martin Maiden & Erich Round. 2019. Romance verbal inflection dataset 2.0. doi:10.5281/zenodo.3552367.
- Bonami, Olivier, Gauthier Caron & Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer & Sophie Prévost (eds.), *Actes du quatrième congrès mondial de linguistique française*, 2583–2596.
- Carroll, Stephanie Russo et al. 2020. The CARE principles for indigenous data governance. *Data Science Journal* 19. doi:10.5334/dsj-2020-043.
- Corbett, Greville G. 2013. Paradigm conventions. Paper at the 46th Annual Meeting of the Societas Linguistica Europaea, Split, Croatia. 18-21 September 2013.
- Farrar, Scott & D Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT international* 7(3). 97–100.
- Feist, Timothy & Enrique L. Palancar. 2015. Oto-Manguean Inflectional Class Database. University of Surrey. doi:10.15126/SMG.28/1.
- Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5. 180205. doi:10.1038/sdata.2018.205.
- Fowler, Dan, Jo Barratt & Paul Walsh. 2018. Frictionless data: Making research data quality visible. *International Journal of Digital Curation* 12(2). 274–285. doi:10.2218/ijdc.v12i2.577.
- Maiden, Martin et al. 2010. Oxford online database of romance verb morphology. Online website. Browsable database. <http://romverbmorph.clp.ox.ac.uk/>.
- Malouf, Robert, Farrell Ackerman & Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the society for computation in linguistics 2020*, 446–456. New York: ACL. <https://aclanthology.org/2020.scil-1.52>.
- McCarthy, Arya D. et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the twelfth language resources and evaluation conference*, 3922–3931. Marseille, France: European Language Resources Association. <https://aclanthology.org/2020.lrec-1.483>.
- McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar & Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, 19–21.
- Pellegrini, Matteo & Marco Passarotti. 2018. LatInfLexi: an Inflected Lexicon of Latin Verbs. In Elena Cabrio, Alessandro Mazzei & Fabio Tamburini (eds.), *Proceedings of the fifth italian conference on computational linguistics (clic-it 2018)*, vol. 2253 CEUR Workshop Proceedings, December. <http://ceur-ws.org/Vol-2253/paper23.pdf>.
- Stump, Gregory T. & Raphael Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge: Cambridge University Press.
- Thornton, Anna M. 2012. Reduction and maintenance of overabundance. a case study on italian verb paradigms. *Word Structure* 5(2). 183–207.
- Wilkinson, Mark D. et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3(1). 160018. doi:10.1038/sdata.2016.18.

# Disentangling morphomic splits in Limbu

*Berthold Crismann*      *Baptiste Loreau Unger*  
LLF, CNRS & U Paris Cité    ENS Paris-Saclay & U Paris Cité

In this talk, we shall examine patterns of syncretism in the system of participant marking in Limbu (van Driem, 1987), a Kiranti language spoken in eastern Nepal by 180,000 people. Similar to other Kiranti languages, such as Athpare (Ebert, 1997b) or Camling (Ebert, 1997a), Limbu verbs inflect for their core arguments, corresponding to S(ole), A(gent) and P(atient) roles. Participants are marked for number (singular, dual, plural) and person (1,2,3), including an inclusive/exclusive distinction for first person non-singular. Marking is predominantly suffixal, with only a few prefixal markers for number, person, and negation.

On the one hand, participant marking in Limbu appears relatively transparent: in case of combination, participants tend to be marked individually, although cases of portmanteau marking do exist. Adding to the transparency, person and number distinctions for each participant are often marked separately by discrete markers. Moreover, the system of participant marking is largely the same across different tenses (non-past vs. past) or polarity.

On the other hand, this transparency contrasts with a number of syncretism patterns that affect different parts of the paradigm in different ways (cf. Table 1 for reference). Besides almost complete neutralisation of second person number contrasts in the  $2 > 1$  and  $1 > 2$  cells<sup>1</sup> (see §1.3), we also find partial neutralisation of the dual/plural distinction for third person A and third person P (see §1.2). This syncretism differs for A and P roles, providing an instance of divergent bidirectional syncretism in the terminology of Stump (2001). The third type of syncretism that complicates the system can be observed with allomorphic variation of person/number markers in different tenses (or polarities), giving rise to what we shall call “pseudo-Paninian” splits (see §1.1).

Table 1: Limbu person marking<sup>2</sup> (based on conjugation lists in van Driem, 1987, 368-374)

↓ A \ P →	1SG	1DE	1PE	1DI	1PI	2SG	2DU	2PL	3SG	3DU	3PL				
1SG						-nɛ	-nɛ-tchi-ŋ	-n-i-ŋ	-u-ŋ   $\frac{-ʔɛ}{-paŋ}$	-u-ŋ-si-ŋ   $\frac{-ʔɛ-n-chi-n}{-paŋ-si-ŋ}$	→				
1DE						←	-nɛ-tchi-ge	→	-s-u-ge	-s-u-si-ge	→				
1PE						↙	↓	↘	$\frac{-u-m-be}{-mʔna}$	$\frac{-u-m-si-m-be}{-mʔna-si}$	→				
1DI													a- -s-u	a- -s-u-si	→
1PI													a- -u-m	a- -u-m-si-m	→
2SG	$\frac{kɛ- -ʔɛ}{kɛ- -aŋ}$	↑	↗						kɛ- -u	kɛ- -u-si	→				
2DU	←	a-ge-	→						kɛ- -s-u	kɛ- -s-u-si	→				
2PL	↙	↓	↘						kɛ- -u-m	kɛ- -u-m-si-m	→				
3SG	$\frac{-ʔɛ}{-aŋ}$	-si-ge	-i-ge	a- -si	a-	kɛ-	kɛ- -si	kɛ- -i	-u	-u-si	→				
3DU	↑	↑	↑	↑	↑	↑	↑	↑	-s-u	-s-u-si	→				
3PL	$\frac{mɛ- -ʔɛ}{mɛ- -aŋ}$	mɛ- -si-ge	mɛ- -i-ge	a-m- -si	a-m-	kɛ-m-	kɛ-m- -si	kɛ-m- -i	mɛ- -u	mɛ- -u-si	→				
S →	$\frac{-ʔɛ}{-aŋ   -paŋ, -aŋ}$	-si-ge	$\frac{-i-ge}{-mʔna}$	a- -si	a-	kɛ-	kɛ- -si	kɛ- -i	-0	-si	mɛ-				

<sup>1</sup>We use the (standard)  $m > n$  notation to denote any participant with person/number features  $m$  acting on a patient with features  $n$ .

<sup>2</sup>Cells with allomorphy conditioned by tense or polarity are divided up into 2 by 2 subtables, with non-past at the top, past at the bottom, affirmative on the left and negative on the right. Regular tense and polarity marking has been omitted from the paradigm in the interest of readability (largely reproducing the non-past affirmative paradigm).

We use arrows to represent syncretism between adjacent cells, without necessarily implying any directionality.

# 1 Syncretism in Limbu participant marking

## 1.1 “Pseudo-Paninian” splits

The first case of syncretism patterns we shall discuss is witnessed in at least two places in the paradigm in Table 1: one involving the 2 > 1 paradigm (centre left in Table 1), the other involving exponents of first plural exclusive across different tenses and roles (cf. Table 2a).

Let us start with the 2 > 1 case. When taken in isolation, it just looks like your standard Paninian split: the *a-ge-* prefix serves to express all cells of this 3 by 3 sub-paradigm, effectively neutralising number distinctions, while there is a special circumfixal form for the 2SG > 1SG cell that functions as an override. This form itself is peculiar: the *ke-* prefix also serves as a second-person marker notably in the 2 > 3 transitive and the 2 intransitive subparadigms. Similarly, the suffix, which features the two allomorphic variants *-ʔε* (NPST) and *-aŋ* (PST), can also be found in the 3 > 1SG cells and the 1SG cell of the intransitive paradigm. Thus, we are faced with the paradoxical situation that what functions locally as an override in the 2 > 1 subparadigm actually corresponds to more general forms used elsewhere in the expression of second and first person participants.

Table 2: Schematic representation of pseudo-Paninian splits

(a) First plural exclusive							(b) Dual/plural (-si vs. mε-)				
Role	A > 3SG		S		3SG > P		Role	A > 1/2	A > 3	S	P
Tense	NPST	PST	PST	NPST	PST	NPST					
1PE	-u-m-be	-mʔna	-mʔna	-i-ge	-i-ge	-i-ge	3DU	mε-	-si	-si	-si
2PL	ke- -u-m	ke- -u-m	ke- -i	ke- -i	ke- -i	ke- -i	3PL	mε-	mε-	mε-	-si
1PI	a- -u-m	a- -u-m	a- -ε	a-	a- -ε	a-					

A highly similar behaviour can be observed for marking of first plural exclusive: looking at the 1PE column in both the transitive and intransitive paradigms, it appears that *-mʔna* is a specific portmanteau override in the past intransitive paradigm for the otherwise regular *-i-ge*, the latter being composed of the exclusive marker *-ge/-be* and the plural marker *-i*. Both exclusive and plural markers are attested in other areas of the full paradigm as well, such as the first exclusive dual (1DE) and plural (1PE) rows (*-ge/-be*) and the 2PL column (*-i*).

However, if we look more closely at the distribution of *-mʔna*, we find it in the 1PE > 3 cells as well, where it is the past tense allomorph of *-u-m-be*. Again, each of these three markers is fairly general, marking third person P (*-u*), first/second plural A (*-m*), and first person exclusive (*-ge/-be*). Given the syncretism of *-mʔna* across S and A roles, it turns out to be more general than what we expect of a simple Paninian override: while it is indeed more specific than its competitors in most respects, combining past with first person exclusive plural, its role specification, viz. A or S (= “nominative”), is actually not more specific than either *-i* (S or P = “absolutive”) or *-m* (A). The nature of the split is shown schematically in Table 2a: as one can easily discern, the split is neither fully natural (or balanced), nor fully Paninian. However, merely considering it as morphomic misses the clean separation of the (green) *-u-m* and the (red) *-i* cells. Furthermore, if Paninian competition can be invoked, it will be possible to maintain a natural description of the competitors *-u-m-be* and *-i-ge* in terms of the general properties expressed by their constituent formatives.

We shall argue that the issue with pseudo-Paninian splits can be resolved in this case by generalising the shared properties of *-mʔna* into a more abstract common rule type, yet expand this rule type into two rules that are individuated for the specific argument role (S vs. A).

## 1.2 Neutralisation of dual/plural

Although number marking in Limbu generally distinguishes dual and plural, there are regions in the paradigm in Table 1 where this distinction is effectively neutralised: most obviously,

for third person P participants, *-si* functions as a mere non-singular marker. In the  $3 > 1/2$  cells, the dual/plural distinction is equally neutralised for A participants, now featuring *mε-* as the exponent of non-singular third person A. However, in situations where the contrast between dual and plural is maintained, as e.g. for third person intransitive S, *mε-* serves to mark plural, whereas *-si* expresses dual. One way to picture this situation is in terms of divergent bidirectional syncretism (Stump, 2001) where the dual marker takes on expression of plural in the third person P cells, and the plural marker is extended to expression of dual cells in the third person A cells.

As depicted in Table 2b, the syncretism of *-si* and *mε-* in third person gives rise to a pattern of two interlocking L shapes. Thus, when taken in isolation, this looks like a “balanced” morphomic split where neither of the two markers can receive a straightforward natural characterisation, yet none of the two can be considered a default or an override either. The picture changes, however, once we include the full range of exponents for non-singular number: as it turns out, there is no other marker that uniquely encodes dual, but there are other markers (*-m, -i*) that specifically encode plural. As a consequence, it is safe to regard *-s(i)/-tchi* as a non-singular marker that only gets restricted to dual by virtue of (Paninian) competition with a dedicated plural marker. This is in line with the fairly wide distribution of *-si*: according to van Driem (1987), the alternation between *-si*, *-s* and *-tchi* is in most cases a mere phonologically conditioned allomorphy. Under this perspective, *mε-* is an A/S third person non-singular marker with two specialised instances: ambiguity-preserving in  $3 > 1/2$  cells, and plural otherwise. In sum, we can resolve the split similar to the case of *-m?na* discussed in §1.1 above.

### 1.3 Neutralisation of number

A particular neutralisation pattern affects the cells with only speech act participants ( $2 > 1$ ,  $1 > 2$ ). Person marking for  $2 > 1$  uses a combination of the role-independent markers for first (*a-*) and second person (*kε-*) participants. Person marking for  $1 > 2$ , by contrast, is expressed by the portmanteau marker *-nε*, encoding a first person A acting on second person, which preempts the role-independent markers via Panini’s principle. Number however remains entirely unmarked in the  $2 > 1$  cells and for P in the  $1NSG > 2$  cells. These cells are not only clearly exceptional in Limbu, but also in other Kiranti languages such as Athpare or Camling, giving rise to cross-linguistic variation.

## 2 Towards a formal analysis

Previous formal analyses of Limbu participant marking have so far largely focused on the phenomenon of affix copying found with the *-ŋ* and *-m* markers (Zimmermann, 2012; Stump, 2022). Stump does provide a grammar fragment for part of the paradigm, but the intransitive and third person agent sub-paradigms are not covered. Thus, the specific issues of syncretism we are confronted with here have so far not been addressed.

The analysis we propose is couched in terms of Information-based Morphology (= IbM; Crysmann & Bonami, 2016), an inferential-realisation theory of inflection that implements a templatic view of morphotactics within a formalism based on inheritance hierarchies of rules.

To start with, let us consider how the pseudo-Paninian split for *-m?na* can be captured: Figure 1a provides the relevant rules for portmanteau *-m?na* and its competitors *-u-m-be* and *-i-ge*. Crucially, the rules for *-m?na* are organised in a type hierarchy where the supertype generalises across the S and A cells. As can be easily verified, this supertype is neither more general nor more specific than its competitors or their combinations, since it is more informative with respect to tense, but less (*s-or-a* vs. *a*) or incommensurate (*s-or-a* vs. *s-or-p*) with respect to role. By providing subtypes for *-m?na*, however, we can individuate the general constraints to the specific roles (*a* vs. *s*), such that *-m?na* can serve as a true Paninian override in these contexts.

Turning to non-singular marking (cf. Figure 1b for a partial set of rules), there are two issues that need to be solved: first, constrain the macro-distribution of inflectional marking across the

paradigm (cf. §1.3), and second, orchestrate the competition between exponents (cf. §1.2). To address the former, we provide partial rules in the INFL dimension that constrain non-singular marking to two regular areas (expression of S/presence of a third person participant), as well as exceptional marking in the 1 > 2 area. Rules of exponence in the EXPO dimension are cross-classified with these constraints on inflectedness, accounting for the absence of non-singular marking in the 2 > 1 cells and the restricted distribution of *-si* and *-i* in the 1 > 2 cells. With respect to exponence, we regard *-si* as the most general marker of non-singular, since it can be found in all three persons and all three roles (S, A, P). In case *-si* does not surface, there is either a (plural) competitor, or else no expression of number altogether. The rules for *-si* do not by themselves disambiguate between dual and plural, but they do positionally distinguish role (cf. Swahili; Crysmann & Bonami, 2016). Disambiguation of number arises by competition with markers that are either inherently plural (like *-i*, *-m* or plural inclusive  $\emptyset$ ), or are specialised to plural in the relevant cells (*mε-*). Presence of non-singular *mε-* in the 3 > 1/2 cells is equally derived by Panini’s principle, given that the only non-singular competitor (*-si*) is more general, not bearing any person specification.

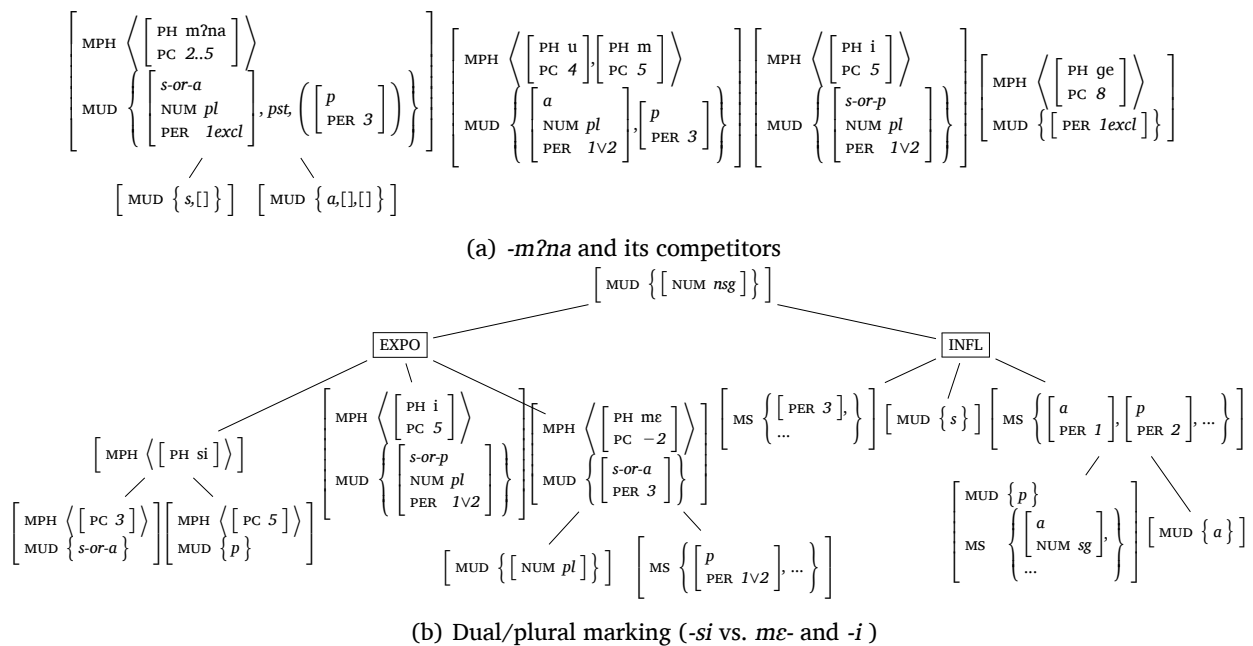


Figure 1: Rule hierarchies

To conclude, pseudo-Paninian splits and exceptional neutralisation of number marking in Limbu highlight the usefulness of underspecification and cross-classification in hierarchies of inflectional rules, to better encode and reconcile conflicting generalisations in complex morphological systems.

## References

- Crysmann, Berthold & Olivier Bonami. 2016. Variable morphotactics in Information-based Morphology. *Journal of Linguistics* 52(2). 311–374.
- van Driem, George. 1987. *Limbu language*. Berlin: Mouton de Guyter.
- Ebert, Karen. 1997a. *Camling (Chamling)*. München/Newcastle: Lincom Europa.
- Ebert, Karen. 1997b. *A grammar of Athpare*. München/Newcastle: Lincom Europa.
- Stump, Gregory T. 2001. *Inflectional morphology. A theory of paradigm structure*. Cambridge: Cambridge University Press.
- Stump, Gregory T. 2022. *Morphotactics: A rule-combining approach*. Cambridge University Press.
- Zimmermann, Eva. 2012. Affix copying in Kiranti. In Enrico Boone, Kathrin Linke & Maartje Schulpen (eds.), *Proceedings of ConSOLE XIX*, 343–367. Leiden: Leiden University.

# From sound change to overabundance: the history of wa-/wu- and mba-/mbu- prefixal allomorphy in Central Pame (Otomanguean)

*Borja Herce*  
University of Zurich

The term 'overabundance' (Thornton 2011) refers to the situation where multiple forms are possible for the expression of the same morphosyntactic feature bundle in a particular lexeme (e.g. dreamed~dreamt 'dream.PST'). The last decade has witnessed an increased interest in the phenomenon, whose typology (Thornton 2019), relationship to other paradigmatic phenomena (e.g. defectivity), and its diachronic sources still need to be investigated further.

I focus on the latter by exploring in detail a striking case of overabundance in Central Pame (cent2145, Otomanguean) which affects various (often high frequency) forms in the paradigm of many lexical items. It involves free choice between the prefixes wa- and wu-, and mba- and mbu-. These occur in various word classes and with various different values. In verbs (see below), they can occur, depending on the inflection class, as markers of 3SG.PRS, 3PL.PRS, and SUB. Grammaticality judgements from my consultants indicate that /a/ and /u/ versions of these prefixes are both acceptable in largely every lexeme and value where they occur.

TAM	Person	'feel'	'die'		'deceive'
PRS	1SG	la-ttsáu?	la-ttũ		tu-nhũn
	1PL.EX	ta-ttsáun?	ta-ttũn?		tu-nhũn?
	2SG	ki-ttfáu?	ki-kjũ		tu-nhũn
	2PL	ki-ttfáun?	ki-kjũn		tu-nhũn
	3SG	<b>wa</b> -ttsáu? <b>wu</b> -ttsáu?	∅-ttũ		lu-nhũn
	3PL	∅-ttsháu?	∅-ttũt		<b>wa</b> -nhũn <b>wu</b> -nhũn
SUB	1SG	nda-ttsáu?	<b>mba</b> -ttũ	<b>mbu</b> -ttũ	ndu-nhũn
	1PL.EX	nda-ttsáun?	<b>mba</b> -ttũn?	<b>mbu</b> -ttũn?	ndu-nhũn?
	2SG	ŋgi-tfáu?	ŋku-ttũ		ŋgi-ŋhjũn
	2PL	ŋgi-tfáuin?	<b>mba</b> -ttũn	<b>mbu</b> -ttũn	ŋgi-ŋhjũn
	3SG	nda-tsáu?	<b>mba</b> -ttũ	<b>mbu</b> -ttũ	nda-nhũn
	3PL	nda-tsháu?	<b>mba</b> -ttũt	<b>mbu</b> -ttũt	nda-nhũn

Table 1: Partial paradigms of three Central Pame verbs

To identify the possible synchronic probabilistic predictors and the diachronic source of this curious case of overabundance, I conducted research of two types:

1) Elicitation: I compiled a list with 254 possibly overabundant word forms (plus an equal number of 'distractors', i.e. non-overabundant words). These were elicited three times (in separate sessions) in a randomized order from two native speakers of different ages to allow for a diachronic interpretation (M29, F45, consider the 'apparent time' studies in Sociolinguistics, see Bailey et al. 1991).

2) Corpus: From the earliest 1950 texts and descriptions of the language (Gibson 1950a, 1950b, 1950c), and a translation of Saint John's Gospel from the 1970's, I mined all the words suspected of being overabundant. In total, 275 word forms were collected from the former period, and 622 from the latter.

Each of the elicited and corpus-mined word tokens were classified according to whether they contained the /a/ or /u/ form of the prefix (the predicted variable) and according to five different possible predictor variables: i) the morphosyntactic value of the form (e.g. 3SG.PRS, 3PL.PRS, SUB, see Table 1), ii) the nature of the following (stem) consonant (e.g. /p/, /t/, /k/, /m/, /n/, etc.), iii) the structure of the syllable (i.e. open, closed) where the prefix is found, iv) the (first) vowel of the following stem (e.g. /a/ in 'feel', /u/ in 'die' in Table 1), and v) whether the prefix is found in a stressed or an unstressed syllable.

	1950			M29		
	Chisq	Df	Pr(>Chi)	Chisq	Df	Pr(>Chi)
Morphosyntactic value	7.598	6	0.49	48.086	6	<.001 ***
Following consonant	65.268	12	<.001 ***	49.470	13	<.001 ***
Syllable structure	0.942	2	0.58	0.198	2	0.91
Stem vowel	1.525	4	0.69	153.647	4	<.001 ***
Stress	6.822	1	<0.01 **	0.642	1	0.42

Table 2: Results of a binary logistic regression model assessing the predictability of the /a/ /u/ allomorphy from different grammatical and phonological factors

Results from a binary logistic regression model (using the function glm in R) reveal that the choice between the /a/ or /u/ forms of the prefix was initially almost perfectly predictable from the phonological environment, particularly from the nature of the following consonant. Looking at these in more detail we find that the /u/ allomorphs are found almost exclusively before a bilabial consonant (38 out of 42 forms with /u/ occur before a bilabial, for only 29 out of 233 forms with /a/). Furthermore, we find that by 1975, the requirement for the /u/ allomorph to occur before bilabials had become almost categorical. This points towards a (regular) sound change BaB>BuB in the earliest periods. The articulatory motivation of such change would be, furthermore, quite clear, constituting an assimilation motivated by gestural dynamics in the typology of sound changes by Garrett & Johnson (2013:19).

Rather than preserving the inherited phonological conditioning of the resulting allomorphy (i.e. /u/ before bilabial, /a/ elsewhere), later generations appear to have turned to different probabilistic predictors. In the elicited data from the youngest contemporary speaker, we observe that the nature of the following stem vowel has become a very important one (with /u/-vowel stems favouring /u/-prefixes), and so has the morphosyntactic value of the form (with SUB and 3PL.PRS preferring /u/ and 3SG.PRS preferring /a/). These emerging trends might be understandable given some pervasive morphophonological processes in the language by which vowels in prefixes can trigger changes in the stem vowel (see Gibson 1956), and given the possibility for morphological analogy to other forms in the paradigm (notice in Table 1 how in the conjugation where wa-/wu- occurs in the 3PL.PRS, all other PRS prefixes contain the vowel /u/, a



factor which might favour this vowel in the 3PL.PRS as well). In addition to the change in predictors, we find that empirically documented overabundance (defined as the proportion of words attested at least twice that appear in both of their possible forms) shows a U-shape in the temporal axis.

	1950	1975	F45	M29
<b>/a/-only</b>	78% (N=46)	64.8% (N=46)	47.3% (N=89)	26.6% (N=50)
<b>overabundant</b>	13.6% (N=8)	1.4% (N=1)	3.2% (N=6)	44.7% (N=84)
<b>/u/-only</b>	8.5% (N=5)	33.8% (N=24)	49.5% (N=93)	28.7% (N=54)

Table 3: Proportion of attested overabundant forms in different periods/speakers

In 1950, most of the words that are attested with /u/ are also attested with /a/. I interpret this early overabundance as the result of sound change in progress during the initial emergence of the /u/ allomorphs (note how these are initially a clear minority). In later years, and in the speech performance of the older present-day speaker (F45), overabundance practically disappears, with most words attested consistently with either a wa-/mba- or a wu-/mbu- prefix. At the most recent stage, however, overabundance makes a swift comeback: /a/-prefixes are reintroduced in the phonological environment they were expelled from through sound change (thus signalling the end of the sound change as a synchronically applicable rule), and other probabilistic predictors (e.g. morphosyntactic values) start to play an important albeit nondeterministic role in the /a/~ /u/ allomorphy.

I believe the explanation for this comeback of overabundance must be sought in paradigm structure. A categorical division between verbs taking wa-/mba- vs wu-/mbu- multiplied by two the number of inflection classes in the language. Furthermore, while morphological distinctions were quite robust between the original inflection classes, the newly-created inflection classes were identical across most of the paradigm. This complicated the Paradigm Cell Filling Problem (PCFP, see Ackerman et al. 2009) quite substantially. Some (younger) speakers appear to have solved this through free choice. If every lexical item is allowed to use wa- and wu-, or mba- and mbu- indistinctly, this brings about a merger of the novel minimally-distinct inflection classes, and a return to the original conjugational system with mostly robust/canonical inflection class distinctions (Corbett 2009).

In conclusion, we find that this case of overabundance in Central Pame opens a fascinating window into the possible diachronic origins and progression of the phenomenon. Bringing up (quantitative) evidence from a non-WEIRD (Henrich et al. 2010) language is especially important here because WEIRD languages have monopolized research on overabundance to date (note that societal differences could very plausibly be associated with empirical differences, e.g. through different degrees of linguistic prescription and standardization). Beyond overabundance, the diachronic developments that have been described here reveal the great power of paradigmatic structure to fight off the disruptive morphological effects of regular sound changes (Sturtevant 1947). Because of the extraordinary importance and pervasiveness of paradigmatic structure in Central Pame, the regularity and original environment of the sound change BaB>BuB have been completely overturned in a very short period of time. This reminds us of the potential/need for paradigms to contribute to the exploration of phylogenetic relations (Meillet 1958, Nichols 1996, Hecce & Bickel forthcoming), particularly in morphology-heavy families where regular correspondences might/should be harder to find.

## References

- Ackerman, Farrell, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins and Juliette Blevins (Eds.), *Analogy in grammar: Form and acquisition*: 54–82. Oxford: Oxford University Press.
- Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand. 1991. The apparent time construct. *Language variation and change* 3, no. 3: 241-264.
- Corbett, Greville G. 2009. Canonical inflectional classes. In *Selected proceedings of the 6th Décebrettes: Morphology in Bordeaux*: 1-11.
- Garrett, Andrew & Keith Johnson. 2013. Phonetic bias in sound change. *Origins of sound change: Approaches to phonologization* 1: 51-97.
- Gibson, Lorna F. 1950a. Three Chichimeca texts. Unpublished manuscript. Online at <https://www.sil.org/resources/archives/57422>.
- Gibson, Lorna F. 1950b. A pedagogical grammar of Central Pame. Unpublished manuscript. Online at <https://www.sil.org/resources/archives/57479>.
- Gibson, Lorna F. 1950c. Verb paradigms in Pame. Unpublished manuscript. Online at <https://www.sil.org/resources/archives/53023>.
- Gibson, Lorna F. 1956. Pame (Otomi) phonemics and morphophonemics. *International Journal of American Linguistics* 22, 4: 242-265.
- Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, no. 7302: 29-29.
- Herce, Borja & Balthasar Bickel. Forthcoming. Paradigmatic predictability metrics as signals of phylogenetic relatedness: a proof of concept in Romance and Pamean diachrony.
- Meillet, Antoine. 1958. *Linguistique historique et linguistique générale*. Société Linguistique de Paris, Collection Linguistique, 8. Librairie Honoré Champion, Paris.
- Nichols, Johanna. 1996. The Comparative Method as heuristic. In Mark Durie & Malcolm Ross (Eds.), *The comparative method reviewed: Regularity and irregularity in language change*: 39-71.
- Sturtevant, Edgar H. 1947. *An Introduction to Linguistic Science*. New Haven: Yale University Press.
- Thornton, Anna. 2011. Overabundance (multiple forms realizing the same cell): a non-canonical phenomenon in Italian verb morphology. In Maiden, Martin et al. (eds.), *Morphological autonomy*, 358-381. Oxford: OUP.
- Thornton, Anna. 2019. Overabundance: a canonical typology. In Rainer, Franz et al. (eds.), *Competition in inflection and word-formation*, 223-258. *Studies in Morphology* 5. Dordrecht: Springer.

---

# Compounding in the Slot Structure Model

Carlos Benavides

University of Massachusetts Dartmouth

---

## 1 Introduction

An account of the semantics of compounding has been one of the most elusive undertakings in morphological research. As Jackendoff (2010) points out, scholars have despaired at finding the range of possible relations (or semantic functions) between the constituents of a compound. The current paper presents a fully developed model of compound formation, set within the framework of the Slot Structure Model (SSM) (Benavides 2003, 2009, 2010, 2022), a constraint-based model of morphology that is based on percolation of both syntactic and semantic features and on slot structure, which organizes the information in the lexical entries of words and affixes. The SSM is partly based on the dual-route model (Pinker 2006, Pinker 1999, Pinker & Ullman 2002). The goal of the paper is to demonstrate how the meaning of a compound is built from that of its constituents, and the relations between them, using the SSM framework.

It is shown that analyzing compound formation using SSM brings with it several advantages, including a more comprehensive explanation of how the semantics of compounding works; a principled, more systematic way to determine the headedness of a compound, regardless of the language; the ability to explain the generativity of compounds on the basis of the actual and potential information contained in the lexical entries of the constituents; and the simplification of the interpretation of compounds, not only because of the notation, but also due to the structure of the lexical entries involved in the determination of compound meaning. Importantly, SSM achieves all this employing the same machinery that is already used for derivation, with some enhancements, including the enrichment of lexical entries, to produce a flexible, generative mechanism that accounts for the semantics of a wide range of compounds types. These include NN, NA, AN, VN, and AA compounds. The analysis is based on English, Spanish and German compounds, but it should be applicable to compounds in other languages. The paper thus achieves a wider coverage of the data than other current approaches that deal with the semantics of compounding, including Jackendoff (2009, 2010, 2016) and Toquero-Pérez (2020), who restrict their analysis to NN compounds, and Schlücker (2016), who discusses AN compounds.

According to Jackendoff (2010), the class of possible meaning relations between the two nouns in a compound is the product of a generative system. This paper shows how the lexical entries of the two constituents of a compound provide the basic information that gives rise to the generativity of compound meaning. An indefinite number of semantic

functions can be generated based on the lexical information of the compound constituents. The unification of the two lexical entries contributes to making it a generative process.

Example compounds to support the analysis have been obtained from the Corpus del Español (CDE, Davies 2016), the iWeb corpus (Davies 2018), Jackendoff (2010), Toquero-Pérez (2020), Lang (2013), Moyna (2011), and Schlücker (2016).

## 2 The Slot Structure Model (SSM)

The SSM is an approach to morphology based in part on Lexical Conceptual Structure (LCS) (Jackendoff 1990, 2002, Rappaport & Levin 1988, 1992) that explains the process of [base + affix] unification in regular word formation in Spanish (e.g. *demoli + cion* [*demolición* ‘demolition’]) and other languages, and is crucially based on the notion of lexical entries instantiated in a slot structure. Employing the mechanisms of subcategorization/selection (subcat/select) and percolation, already available in the generative framework (cf. Lieber 1992, 1998, Pinker 2006, Pinker 1999, Pinker & Ullman 2002, Huang & Pinker 2010), the model unifies all the processes that take place during the formation of a complex word (e.g. *plega + ble* [*fold + able*] ‘foldable’).

Crucial to the SSM is that percolation, subcat/select, and slot structure, acting in concert determine the structure and content of the lexical entries of derivatives and allow for predictions to be made about the behavior of groups of features in the formation of a word. Percolation in particular, as shown by Pinker (1999) and Pinker & Ullman (2002), is key to account for compositionality in word formation. Huang & Pinker (2010) call percolation *information-inheritance* and stress the need for this mechanism in morphology, both in inflection and word formation.

In addition to accounting for regular derivation, the SSM adequately accounts for regular inflection (e.g. *libro + s* ‘book + s’, *beb + o* [drink-1sg, pres.] ‘I drink’), as well as the regular derivational morphology of several languages genetically unrelated to Spanish (Mam, Turkish, Swahili). In addition, the SSM has been extended (Benavides 2003, 2009, 2022), using the exact same tools and mechanisms, to other types of affixes (in Spanish and other languages), namely, derivational prefixes, passives, expressive suffixes (e.g. diminutives), inflectional affixes, and parasynthetics, as well as to causatives and applicatives in Chichewa, Madurese, Malayalam, Chimwi:ni, and Choctaw. This suggests that the notions of percolation, subcat/select, slot structure and the LCS may be universal constructs.

Diagrams presented in the paper demonstrate how the semantics of compound formation is implemented with an adapted SSM formalism, and show that predictions can be made about the organization of information, including argument structure, in the resulting compound. For example, Diagram 1 shows the formation of the compound *plastic bag*. Each

column represents a lexical item, with its respective slots, and the arrows indicate that a feature from *plastic* has percolated to the COMPOSITION (COMP) slot of the entry for *bag*, the head of the compound, resulting in *plastic bag* as an item with a unified meaning.

Diagram 1

plastic	bag
<u>CATEGORIAL</u> [THING] N	<u>CATEGORIAL</u> [THING] N
	<u>CORE</u> ARTIFACT BAG
	<u>PF</u> HOLD CONTENT
<u>CORE</u> MATERIAL PLASTIC →	<u>COMP</u> → PLASTIC

This type of representation has an advantage over Jackendoff's (2009, 2010, 2016) functions (e.g. COMP ( $X_1, Y_2$ ) 'N<sub>2</sub> is composed of N<sub>1</sub>') in that it enables the basic functions to be integrated into the lexical entries of the constituents, thus allowing for an easier interpretation. It also allows for more accurate predictions to be made about the meaning of compounds, because information inside the lexical entries of the constituents compose with each other inside the entries.

### 3 Conclusion

This paper shows that an analysis of compounding that employs the SSM framework brings about several important advantages, as outlined in §1. This is the case because the information related to the semantic functions is shown in the context of the rest of the semantic information of the lexical entries of the compound constituents. Importantly, all this is accomplished with the same machinery that is already used for derivation. The key innovation of the model is the enrichment of lexical entries to produce a flexible, generative mechanism that accounts for the semantics of a wide range of compounds. The generativity comes from the pieces of information inside the lexical entries of the constituents, which interact with pragmatics and compose with each other inside the entries, not detached from them as in Jackendoff (2009, 2010, 2016), Toquero-Pérez (2020) and Schlücker (2016).

## References

- Benavides, Carlos. 2003. Lexical Conceptual Structure and Spanish derivation. *Journal of Language and Linguistics* 2. 163-211.
- Benavides, Carlos. 2009. *The semantics of Spanish morphology: The Slot Structure Model*. Saarbrücken: VDM Verlag.
- Benavides, Carlos. 2010. El clítico 'se' en español y la estructura léxico-conceptual. *RILCE:Revista de Filología Hispánica* 26. 261-88.
- Benavides, Carlos. 2022. Morphology Within the Parallel Architecture Framework: The Centrality of the Lexicon Below the Word Level. *Isogloss. Journal of Romance Linguistics* 8(1)/7. 1-87. DOI: <https://doi.org/10.5565/rev/isogloss.200>
- Davies, Mark. 2016. *Corpus del Español*. <https://www.corpusdelespanol.org>.
- Davies, Mark. 2018. *The iWeb Corpus*. <https://www.english-corpora.org/iweb>.
- Huang, Yi Ting, & Steven Pinker. 2010. Lexical semantics and irregular inflection. *Language and Cognitive Processes* 25. 1411-61.
- Jackendoff, Ray .1990. *Semantic Structures*. Cambridge, MIT Press.
- Jackendoff, Ray. 2002. *Foundations of Language*. Oxford: Oxford University Press.
- Jackendoff, Ray. 2009. Compounding in the Parallel Architecture and conceptual semantics. In Rochelle Lieber & Pavol Stekauer (eds.), *The Oxford Handbook of Compounding*, 105–129. New York: Oxford University Press.
- Jackendoff, Ray. 2010. The ecology of English noun-noun compounds. In Ray Jackendoff (ed.), *Meaning and the lexicon*, 413-451. Oxford: Oxford University Press.
- Jackendoff, Ray. 2016. English Noun-Noun compounds in conceptual semantics. In Pius ten Hacken (ed.), *The Semantics of Compounding*, 15–37. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9781316163122.002>
- Kageyama, Taro, & Michiaki Saito. 2016. Vocabulary strata and word formation processes. In Taro Kageyama and Hideki Kishimoto (eds.), *Handbook of Japanese Lexicon and Word Formation*, 11-50. Berlin: De Gruyter.
- Lang, M.F. 2013. *Spanish word formation*. London: Taylor and Francis.
- Lieber, Rochelle. 1992. *Deconstructing morphology*. Chicago: The University of Chicago Press.
- Lieber, Rochelle. 1998. The suffix -ize in English: Implications for morphology. In Steven Lapointe, Diane Brentari, & Patrick Farrel (eds.), *Morphology and its relation to phonology and syntax*, 12-33. Stanford, CA: CSLI.
- Lieber, Rochelle. 2019. Theoretical issues in word formation. In Jenny Audring & Francesca Masini (eds.), *The Oxford handbook of morphological theory*, 34-55. Oxford: Oxford University Press.
- Moyna, María Irene. 2011. *Compound words in Spanish: Theory and history*. Amsterdam: John Benjamins. DOI: <https://doi.org/10.1075/cilt.316>
- Pinker, Steven. 1999. *Words and rules*. New York: Basic Books.
- Pinker, Steven. 2006. Whatever happened to the past tense debate? In Eric Bakovic, Junko Ito & John J. McCarthy (eds.), *Wondering at the natural fecundity of things: Essays in Honor of Alan Prince*, 221-38. UC Santa Cruz: Festschrifts.
- Pinker, Steven, & Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences* 6. 456-63.
- Rappaport, Malka, & Beth Levin. 1988. What to do with theta-roles. In W. Wilkins (ed.), *Syntax and semantics, Vol. 21*, 7-36. New York: Academic Press.
- Rappaport, Malka, & Beth Levin. 1992. -er nominals: Implications for the theory of argument structure. In Tim Stowell & Eric Wehrli (eds.), *Syntax and semantics, Vol. 26*, 127-53.

New York: Academic Press.

Schlücker, Barbara. 2016. Adjective-noun compounding in Parallel Architecture. In Pius ten Hacken (ed.), *The Semantics of Compounding*, 178-191. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9781316163122.002>

Toquero-Pérez, Luis Miguel. 2020. The semantics of Spanish compounding: An analysis of NN compounds in the Parallel Architecture. *Glossa* 5. 41.1-31. DOI: <https://doi.org/10.5334/gjgl.901>

---

# Verbal-nexus and attributive-appositive N+N compounds in Italian

## A diachronic study

Jan Radimský  
University of South Bohemia

M. Silvia Micheli  
University of Milano - Bicocca

---

### 1 Introduction

Italian Noun+Noun compounds (NN compounds, henceforth) have been extensively investigated from a synchronic point of view (see, for an overview, Radimský 2015) due to their significant productivity and the wide variety of patterns attested in Contemporary Italian. Over the last two decades, several studies have focused on the classification of this type of compounds (see, among others, Baroni, Guevara & Pirrelli 2009), which includes a rather heterogeneous set of words, as well as on properties of specific subtypes (see, e.g., Grandi 2009; Grandi, Nissim & Tamburini 2011 and Radimský 2016 on the attributive-appositive compounds, or Baroni, Guevara & Zamparelli and Lami & van den Weijer 2022 on verbal-nexus compounds, according to the classification proposed by Scalise & Bisetto 2009).

On the other hand, much less attention has been paid to the diachrony of NN compounds, that seem to represent a relatively recent innovation in Romance. According to Rainer (2021), the pattern does not display any continuity from Latin compounding and rather stems from a variety of heterogeneous syntactic constructions whose number seems extremely limited in Italian, at least until the end of the 19<sup>th</sup> Century.

The aim of this contribution is to examine thoroughly the diachronic profile of two prominent Italian N+N compounding patterns, namely verbal-nexus NNs (such as *ritiro<sub>N</sub> bagagli<sub>N</sub>* – “baggage claim”) and attributive-appositive NNs (such as *parola<sub>N</sub> chiave<sub>N</sub>* – “keyword”), based on a large sample of more than 4.000 manually filtered compounds (types) and their diachronic frequency profiles drawn from the Google n-gram data. With reference to the theoretical frameworks of Construction Morphology (Booij 2010, 2016), Relational Morphology (Jackendoff & Audring 2020) and Diachronic Construction Grammar (Hilpert & Gries 2009, Traugott & Trousdale 2013, Goldberg 2019, Hilpert 2021, among others) we will analyse the progressive coinage of constructions at different levels of abstraction (i.e. substantial, semi-schematic and schematic), the relationship between them and the factors that trigger the “productivity upgrade” of the respective schemas.

Indeed, it is not very often that a new compounding pattern appears and develops in a modern language, in a diachronic period that is quite richly documented by written sources. Therefore, the analysis of this process will not only make it possible to show the specific situation of Italian NN compounds, but also to discuss general theoretical questions concerning the emergence of compounding patterns within the selected framework, such as *coverage* (Goldberg 2019) or *structural intersection* (Jackendoff & Audring 2020), as well as methodological tools designed for analysis of diachronic corpus data, such as *variability-based neighbour clustering* (Hilpert & Gries 2009).

### 2 Verbal-nexus and attributive-appositive NN compounds

#### 2.1 Key properties

Verbal-nexus NNs (henceforth VNX NNs, also referred to as Argumental NNs) and attributive-appositive NNs (henceforth ATAP NNs) represent two prominent patterns of left-headed present-day Italian NNs (see Radimský 2015, among others).

VNX NNs (such as *trasporto<sub>N</sub> merci<sub>N</sub>* – transport of goods) are a subtype of left-headed endocentric subordinate compounds consisting of a deverbal head and a non-head element



which is interpreted as its argument. The interpretation of a VNX NN is triggered by the deverbal head (i.e. the leftmost element), so that these compounds are expected to form head-based ‘families’ or ‘semi-schematic constructions’ (such as *trasporto-N* – N-transport); however both synchronic and diachronic data surprisingly suggest that they also form argument-based families (such as *N-merci* – N-goods) (cf. Radimský 2020, 2023 in press). According to various scholars, Italian VNX NNs represent the most – if not the only really – productive higher-order subordinate NN construction in Romance (Rainer 2016, Baroni, Guevara & Zamparelli 2009, Radimský 2018).

ATAP NNs (such as *parola<sub>N</sub> chiave<sub>N</sub>* – “keyword”) feature a head-modifier attributive relationship that may be paraphrased as ‘N1 is a (kind of) N2’. The modifier may have either a metaphoric (in ‘appositive NNs’) or a literal (in ‘attributive NNs’) interpretation, such as in *parola<sub>N</sub> chiave<sub>N</sub>* (“keyword” – the word is ‘key’, important) and *luogo<sub>N</sub> simbolo<sub>N</sub>* (“symbolic place” – the place is a symbol), respectively. In both cases, the interpretation of ATAP compounds is triggered by the modifier (i.e. the rightmost element) and they tend to form strong modifier-based families, which is why selected modifiers with highest type frequencies have sometimes also be analysed as ‘noun-clad adjectives’ (Grandi, Nissim & Tamburini 2011). It is still debatable whether the ATAP pattern as such represents a productive higher-order construction in contemporary Italian or whether its type frequency growth is rather carried out by a small subset of lower-order semi-schematic constructions. Our current data show that the latter solution is probably more in line with reality.

## 2.2 History of Italian NNs

As Rainer (2021:17) puts it: „the evolution and modern proliferation of NN compounds in the Romance languages, unfortunately, has not yet been studied in detail from a diachronic point of view”.

Single instances of Italian ATAP compounds are already attested in Old Italian. Based on the Cudit corpus, Micheli (2020a:91-93) found 3 ATAP NNs in Old Italian (*pescespada* – swordfish, *pescaporco* – grey triggerfish, *arcamensa* – large cupboard) and 15 ATAP NNs in Middle Italian (Micheli 2020a:145, 152-155), but she assumes that the pattern has reached real productivity and dissemination only since the 21st century (Micheli 2020b, 120).

As for subordinate NNs, the existing studies based on literary Italian do not report cases of such compounds attested before 1950 (Tollemache, 1945; Micheli, 2020a, 2020b), but Rainer (2021:17) notes that they became more frequent in contexts related to commerce and industry already since the 19<sup>th</sup> century. In the journalistic style, first examples are assumed to appear around the 1970s (Dardano 2009:226-229), from where they gradually made their way beyond the narrow sphere of professional communication.

It can be therefore assumed that substantial turning points in the evolution of Italian NN compounds – still very little explored – occurred in the past two centuries.

## 3 Theoretical framework

Construction Morphology and Relational Morphology are usage-based models, which entails that schemas available in the Constructicon capture generalizations over a critical mass of already attested words. In a diachronic perspective thus, “constructionalization” must be based on previous individual “innovation” (in the sense of Traugott & Trousdale 2013). One of the targets of the research is therefore also to find a method for identification of such lexical innovations (or *leader words*) in the early stages of the development of patterns.

Once a critical mass of individual lexical innovations is in place, Constructionalization – within the Relational Morphology framework (Jackendoff & Audring 2020) – consists of two

steps. First, relational links between the existing words must be built through the process of “Structural Intersection”, and then it is necessary to determine whether these new relational schemas are productive. In the case of Compounding, we assume that the Structural Intersection yields primarily semi-schematic constructions, in which chunks of forms (either the leftmost or the rightmost component) are shared. Such a view is consistent with the assumption of Laurie Bauer (2017: 74) that “it is not the N+N pattern of compounding which is productive, but patterns with individual lexemes within that”, as well as with the observation of Franz Rainer (2016:2714) that within Italian N+N compounds, “neologisms tend to follow analogues or series of analogues with the same first or second constituent.” We will show that, surprisingly, both VNX and ATAP NN compounds form N1- and N2-based families, though only some of them achieve higher type frequencies, including higher type frequencies of hapax forms, and can be therefore considered as productive.

If subsequent Constructionalization is to yield some higher-order constructions, these should correspond to areas in which examples encountered so far cluster (cf. the notion of *coverage* by Goldberg 2019: 51-73 and its application to compounds by Hilpert 2015). Our data suggest that these higher-order constructions may or may not correspond to “classes” or “types” of compounds.

## 4 Data & methodology

The research is based on extensive diachronic data drawn from the Google books corpus available in the form of raw frequency lists as the 3rd version of Italian Google n-grams,<sup>1</sup> the size of the underlying Google books corpus is 120,410,089,963 tokens. Data for the extraction of N+N compounds come from pre-treated bigrams and trigrams (to capture compounds with space-separated and hyphen-separated components, respectively) from which a sample of roughly 2.000 ATAP and 2.000 VNX compounds has been extracted. In order to achieve a higher accuracy, most compounds have been checked back manually in Google books and many false positives have been eliminated. For each compound, dated numbers of occurrences in Google books are available from 1850 to the present with a year-by-year precision, which makes it possible to analyse in diachrony not only relative token frequencies of single compounds, but also relative type frequencies of different higher-order constructions, such as head-based or modifier-based families (e.g. N-chiave – “key-N”) and fully schematic constructions, and their interaction.

To identify diachronic trends and draw regression lines, Theil-Sen estimator was used and supplemented with the Mann-Kendall test for significance testing (Python implementation by Hussain & Mahmud 2019). These rank-based non-parametric methods are suitable to test any form of dependence (not only linear), they do not assume a normal distribution of errors and they are not sensible to outliers, which makes them particularly suitable for trend identification of word usage in diachronic corpora (Kovář & Herman 2013). Potential turning points in the evolution of patterns are detected using the Variability-based neighbour clustering method (Hilpert & Gries 2009).

## 5 References

- Baroni, Marco, Guevara, Emiliano & Vito Pirrelli. 2007. NN compounds in Italian. Modelling category induction and analogical extension. *Lingue e linguaggio* 6(2). 263—90.
- Bauer, L. 2017. *Compounds and compounding*. Cambridge: Cambridge University Press.
- Booij, Geert. 2010. *Construction Morphology*. Cambridge: Cambridge University Press.

---

<sup>1</sup> <https://storage.googleapis.com/books/ngrams/books/datasetsv3.html>

- Booij, Geert. 2016. Costruction Morphology. In Andrew Hippisley & Gregory Stump (eds.), *The Cambridge Handbook of Morphology*, 424—448. Cambridge: Cambridge University Press.
- Dardano, Maurizio. 2009. *Costruire parole*. Bologna: Il Mulino.
- Goldberg, Adele E. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton: Princeton University Press.
- Grandi, Nicola. 2009. When Morphology 'Feeds' Syntax: Remarks on Noun > Adjective conversion in Italian Appositive Compounds. In *Selected Proceedings of the 6th Décembrettes*, 111—124. Somerville, MA: Cascadilla Proceedings Project.
- Grandi, Nicola, Nissim, Malvina & Fabio Tamburini, Noun-clad adjectives. On the adjectival status of non-head constituents of Italian attributive compounds. *Lingue e linguaggio* 10(1). 161—76.
- Hilpert, Martin 2015. From hand-carved to computer-based: Noun-participle compounding and the upward strengthening hypothesis. *Cognitive Linguistics* 26(1). 1—36.
- Hilpert, Martin. 2021. *Ten lectures on diachronic construction grammar*. Brill. 10.1163/9789004446793
- Hilpert, Martin & Gries, Stefan T. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24(4). 385—401.
- Hussain, Manjurul & Ishtiaq Mahmud. 2019. pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software* 4(39). 1556. <https://doi.org/10.21105/joss.01556>
- Jackendoff, Ray & Jenny Audring. 2020. *The texture of the lexicon: relational morphology and the parallel architecture*. Oxford: Oxford University Press.
- Herman, Ondrej & Vojtech Kovár. 2013. Methods for Detection of Word Usage over Time. In *RASLAN - Seventh Workshop on Recent Advances in Slavonic Natural Language Processing*, 79-85. ISBN 978-80-263-0520-0.
- Lami, Irene & Joost van de Weijer. 2022. Compound-internal anaphora: evidence from acceptability judgements on Italian argumental compounds. *Morphology* 32. 1—30.
- Micheli, M. Silvia. 2020a. *Composizione italiana in diacronia. Le parole composte dell'italiano nel quadro della Morfologia delle Costruzioni*. Berlin/New York: De Gruyter.
- Micheli, M. Silvia. 2020b. *La formazione delle parole. Italiano e altre lingue*. Roma: Carocci.
- Radimský, Jan. 2015. *Noun+Noun compounds in Italian: A corpus-based study*. České Budějovice: Jihočeská univerzita. Edice Epistémé.
- Radimský, Jan. 2016. I composti NN attributivi nel corpus ItWac. In Annibale Elia, Claudio Iacobini & Miriam Voghera, (eds.), *Livelli di analisi e fenomeni di interfaccia. Atti del XLVII Congresso di studi della Società di Linguistica Italiana*, 189—204. Roma: Bulzoni.
- Radimský, Jan 2018. Inflection of binominal ATAP compounds in French and Italian: a paradigmatic account. *Lingue e linguaggio* 17(2). 261—272.
- Radimský, Jan 2020. A paradigmatic approach to compounding. In Jesús Fernández-Domínguez, Alexandra Bagasheva & Cristina Lara Clares (eds.), *Paradigmatic relations in word formation*, 164—185. Leiden: Brill.
- Radimský, Jan. In press. Where did the Italian N+N compounds come from? In Jenny Audring et al. (eds.), *Mediterranean Morphology Meeting* 13. Patras, University of Patras. ISSN: 1826-7491.
- Rainer, Franz. 2016. Italian. In Peter O. Müller, Ingeborg Ohnheiser & Susan Olsen (eds.), *Word-formation. An International Handbook of the Languages of Europe*, 2712—2731. Berlin/Boston: Walter de Gruyter.
- Rainer, Franz. 2021. Compounding: from Latin to Romance. In *Oxford Research Encyclopedia of Linguistics*. <https://doi.org/10.1093/acrefore/9780199384655.013.691>
- Scalise, Sergio & Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber & Pavol Šteckauer (eds.), *The Handbook of Compounding*, 49—82. Oxford: Oxford University Press, Oxford.
- Tollemache, Federico. 1945. *Le parole composte nella lingua italiana*. Roma: Edizioni Roes di Nicola Ruffolo.
- Traugott, Elizabeth C. & Graeme Trousdale. 2013. *Constructionalization and constructional changes*. Oxford: Oxford University Press.

---

# The Interplay of Morpho-Phonology and Semantics in the Processing of Plural Subject-Verb-Number Agreement with Collective Noun Constructions

Kalle Glauch

Ruhr-Universität Bochum

---

## 1 Introduction

This study aims at comparing two major models of dependency formation, the *cue-based retrieval* model (Lewis & Vasishth 2005) and the *marking and morphing* (Bock et al. 2001) model, respectively focusing on number retrieval and number representation. The models main assumptions are tested against the influence of different modifier-types on the processing of *conceptual* (plural) subject-verb number agreement with collective noun constructions in German.

The cue-based retrieval model posits that the agreement-target sets retrieval cues during dependency formation, initiating a process of spreading activation to all items in working-memory whose specifications at least partly match the retrieval cue. For subject-verb agreement in language processing, a plural verb sets the retrieval cues +nominative and +plural, spreading activation to all items with features matching the retrieval cues. Typically, the retrieval process enables accessing the subject head-noun leading to successful subject-verb agreement even in the presence of a local distractor noun in the same DP.

The marking and morphing paradigm suggests that a subject-DP's number valuation (SAP-value) is continuous and determined by the interaction between the referent's notional number, the subject-head noun's number morphology and the number specification of local nouns. Successful agreement is established if the subject's SAP-value during dependency formation aligns with the verb's number specification.

In German, plural subject-verb-number agreement can occur between collective nouns that show a discrepancy between notional plurality and grammatical singularity although it is mostly limited to collectives like *Vielzahl* that denote numerosity and cooccur with modifiers like *der Schüler* to form a collective construction (Löbel 2012).

- (1) [Eine Vielzahl [Fem, Nom, Sg.] der Lehrer]      haben [Pl.] Bier getrunken.  
      'A multitude                                      of the teachers      have              beer drunk.'

Compared to English, German allows for a variety of different modifier-types in collective constructions that can be described by the binary factors  $\pm$ Prepositional and  $\pm$ Definiteness, resulting in four different constructions (Tab. 1).

[ + DEF, -PP]	[ + DEF, + PP]	[ - DEF, -PP]	[ - DEF, + PP]
DPNom[Eine Vielzahl DPGen[der Lehrer]]	DPNom[Eine Vielzahl <sub>PP</sub> [von DPDat[den Lehrern]]]	DPNom[Eine Vielzahl DPGen[Ø Lehrer]]	DPNom[Eine Vielzahl <sub>PP</sub> [von DPDat[Ø Lehrern]]]

Table 1: Modifier-Types for Collective Constructions in German.

Although all of these modifications are mostly interchangeable and can be found in analogous contexts in corpora, I assume that the conceptualization differs slightly depending on the modifier type, following Goldberg's principle of no synonymy (1995).

Specifically, the degree of *partitivity*, i.e. the salience of the superset-implicature triggered by the modifier indicating that the main predication of the sentence does not hold for all entities denoted by the modifier-phrase of the collective construction is assumed to be higher when the variables  $\pm$ Definiteness and  $\pm$ PP take positive values (Lindauer 1995).

## 2 Study 1: Modifier-Type and Partitivity

The current study examined whether the theoretically postulated differences in the degree of partitivity are cognitively real, using a probability-judgement study. In a 2x2 repeated measures design with the binary factors  $\pm$ PP and  $\pm$ DEF, the participants were orally presented with sentences containing collective constructions (2) manipulated by modifier type. Participants were instructed to determine the probability of the existence of other entities as those denoted by the collective construction but not referred to by it.

- (2) Eine Vielzahl der pinken Lehrer fliegt über dem Dorf.  
,A multitude of (the) pink teachers fly above the village.'

48 participants, prescreened for German as native language, were recruited via Prolific. Each participant provided responses to four items per condition.

The results (Fig. 1) provide evidence that supports the hypotheses. The data was analyzed using a linear mixed-effects model with random intercepts for participants, predicting the probability rating as a function of the factors  $\pm$ PP,  $\pm$ DEF and their interaction. The model indicates significant main effects  $\pm$ PP and  $\pm$ DEF ( $p = 0.001$ ;  $p = >0.001$ ), with positive values increasing partitivity.

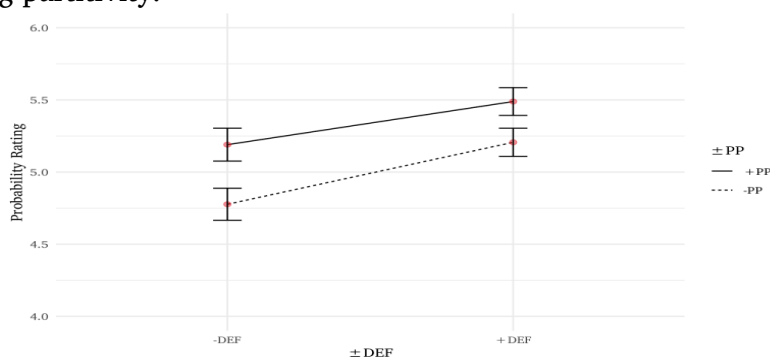


Figure 1: Interaction-plot for partitivity ratings by modifier-type

Following Brehm & Bock (2013), I assume that notionally more coherent sets of entities are perceived as less plural. Furthermore, a higher degree of partitivity of the modifier phrase decreases the coherence of the set of entities denoted by the collective construction due to the contrast with the not-included superset, increasing the notional plurality of the subject-denotation. When processing a sentence with a collective noun construction as the subject, the marking and morphing account predicts an increase in notional plurality to increase the competition between the singular- and plural values in the subject's number representation, increasing number ambiguity as reflected in differences in agreement-processing. The cue-based retrieval model on the other hand does predict any differences purely based on the notional plurality of the modifier as it is not considered relevant in the account.

The notional meaning component of the modifier type can, however, not be viewed in isolation as it is embedded in specific constructions that differ in terms of morpho-phonological factors. Specifically, the case-syncretism of the local noun may be considered relevant in this regard. While the embedding in the dative-controlling preposition *von* (+ PP) causes the local noun *Lehrern* to be marked explicitly as not-nominative, genitive modifiers (-PP) contain case-syncretic local nouns *Lehrer* that have the same form as a nominative noun.

## 3 Study 2: Processing of Plural Agreement by Modifier-Type

According to the marking and morphing account, plural local nouns spread their plural number feature regardless of their morphological form, increasing the overall plural valuation

of the subject. The marking and morphing account consequently predicts the ambiguity in the number valuation and differences in the process of dependency formation to be solely determined by the notional meaning aspect of the modifier phrase. The local noun specified for plural will spread its features regardless of whether it is a noun like *Lehrern* (+PP), explicitly marked for dative case, or a noun like *Lehrer* (–PP) that is syncretic between nominative- and genitive case. The marking and morphing model consequently predicts differences caused by main-effects of  $\pm$ PP and  $\pm$ DEF.

Cue-based retrieval models assume that the plural local noun may be misretrieved during dependency formation due to a partial feature match of +plural with the retrieval cues set by the plural verb. Infelicitous retrieval is, however, significantly more likely if the local noun is case-syncretic to a nominative form as processing the syncretic form will activate the nominative function to some degree so that the local noun seemingly matches both retrieval cues of the verb. Differences in processing are consequently purely determined by the presence of a case-syncretism in the local noun. The cue-based retrieval consequently predicts differences caused by a main-effect  $\pm$ PP.

The study examines the hypotheses about different processing patterns using a speeded-grammaticality judgement procedure with rapid serial visual word presentation. Each word is sequentially displayed for 425ms before automatically disappearing. The design follows a 2x2 repeated measure design with the binary factors  $\pm$ PP and  $\pm$ DEF in the modifier as the manipulated variables. Each sentence consists of a matrix-clause, embedding a subordinate clause with a sentence-final plural verb establishing plural-agreement and a collective construction as its subject.

- (3) Peter | weiß, | dass | eine | Vielzahl | der | Lehrer | in | der | Pause | Bier | trinken.  
'Peter | knows | that | a | multitude | of (the) | teachers | during | break | drink | beer.'

The participant's task is to judge as quickly as possible whether the sentence is grammatical after the last word is presented. Both the judgements and the reaction times are measured as response variables. 70 participants, prescreened for German as native language, were recruited via Prolific. Each participant provided responses to four items per condition.

The assumed main effect for  $\pm$ PP and  $\pm$ DEF from the marking and morphing account is predicted to reflect in more grammatical-judgements as well as longer reaction times when the factors  $\pm$ PP and  $\pm$ DEF take positive values. As notional plurality rises, the competition between singular- and plural number in the subject's SAP-value intensifies, increasing the probability of plural-agreement appearing grammatical. Concurrently, the intensified competition extends the time needed for number feature selection during dependency formation, increasing reaction time.

The cue-based retrieval model, on the other hand, predicts more grammatical-judgements as well as longer reaction times if the factor  $\pm$ PP takes a negative value and the modifier contains a case-syncretic local noun. For reaction times the spread of activation to items matching the verb's +nominative and +plural cue initiated during the retrieval process is distributed between the head noun and the local noun, increasing the retrieval time due to slower activation for either. Acceptance of plural agreement is additive and expected if either the conceptual number specification of the collective head noun – which in contrast to the notional plurality of the local noun can be assumed to influence agreement – is more strongly activated than the singular grammatical number feature during dependency is retrieved, or if the plural local noun is mistakenly retrieved creating an illusion of grammaticality.

The reaction time data (Fig. 2) was analysed using a linear mixed-effects model with random intercepts for participants, predicting the reaction time as a function of the factors

$\pm$  PP,  $\pm$  DEF and their interaction. The model indicates a significant main-effect for  $\pm$  PP ( $p = 0.013$ ), a significant main effect for  $\pm$  DEF ( $p = <0.001$ ) as well as a significant interaction effect ( $p = 0.004$ ). The Grammaticality-Judgement data (Fig. 3) was analysed using a generalized linear mixed-effects model with random intercepts for participants, predicting the probability of grammatical-judgements as a function of the factors  $\pm$  PP,  $\pm$  DEF and their interaction. The model indicates a significant main-effect of the factor  $\pm$  DEF ( $p = 0.017$ ).

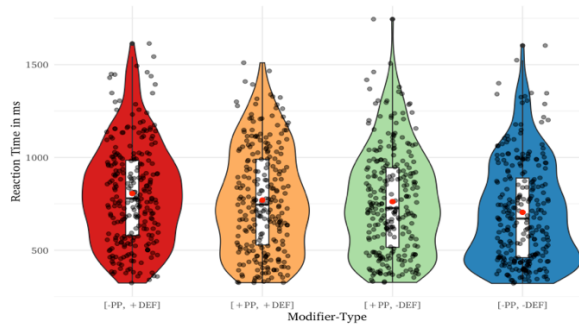


Figure 2: Violin-plot for reaction times

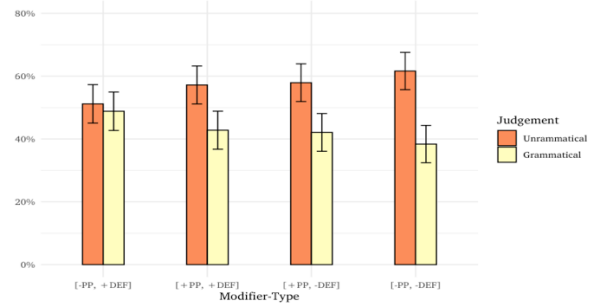


Figure 3: Bar-plot for grammaticality judgements

## 4 Discussion

The observed data cannot be fully explained by either the marking and morphing account nor by the cue-based retrieval account. The prediction of the cue-based retrieval account that only the morpho-phonological aspect of case-syncretism in the modifier influences agreement-processing can neither explain the increased reaction times for the positive value of the factor  $\pm$  PP for the reaction time, nor the increased reaction times and grammatical-judgements for the positive value of the factor  $\pm$  DEF. The predictions of the marking and morphing account fit the data better. Consistent with the predictions, a significant increase in reaction time is observed when the factors  $\pm$  PP and + DEF assume positive values. The factor  $\pm$  PP, however, does not significantly influence the grammaticality-judgements, contrary to expectations. The interaction effect in the reaction times can further not be straightforwardly explained by the marking and morphing model. Inclusion of additional parameters and some combination of the two models, increasing their flexibility, as proposed by Yadav et al. (2023), may yield a superior fit for the data.

## References

- Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, 43(2), 83–128.
- Brehm, Laurel and Kathryn Bock. (2013). What counts in grammatical number agreement? *Cognition* 128: 149-169.
- Corbett, G. Greville. (2000). *Number*. Cambridge: Cambridge University Press.
- Goldberg, Adele E. (1995). *A Construction Grammar Approach to Argument Structure*. The University Press of Chicago.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375–419.
- Lindauer, Thomas. (1995). *Genitivattribute Eine morphosyntaktische Untersuchung zum deutschen DP/NP-System*. Max Niemeyer Verlag, Tübingen.
- Löbel, Elisabeth. (2012). Semantische Kongruenz. In: Härtl, Holden eds. *Interfaces of Morphology*. Akademie Verlag. 201-215.
- Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*.

# Leveling among Patterns of Prosodic Structures of Paradigms for Affix Allomorphy

Koga Hiroki  
Saga University at Saga, Japan

This paper addresses the nonpast affix allomorphy of the Ariake Saga dialect of Japanese, spoken in the area near the coast of the Ariake Sea, and proposes an account using the framework of Optimality Theory by adopting Zodak and Bat-El's (2015) proposal of similarity in leveling. The account predicts and explains leveling and affix allomorphy, that is, the restriction of morphologically well-formed forms to prosodically well-formed ones.

## 1 Data

There are three nonpast affix allomorphs; two are the default and one is the alternative. There are four verb stem types, and the question is which allomorph pairs with which stem:

- (1) Affix allomorphs and stem types
  - a. Nonpast affix allomorphs;  $-(r)u$  (default),  $-uru$  (alternative) (Koga, 2023)
  - b. Verb stem types: C-final (e.g., *tor* 'take', *kir* 'cut', *kaer* 'go home', *sur* 'rub')  
(X)e/(X) (e.g., *tabe/tab* 'eat', *kae/ka* 'change')  
V-final (e.g., *oki* 'get up', *ki* 'wear')  
C/CV (*k/ko* 'come', *s/se* 'do')

## 2 Relevant Accounts

The constraints of prosodic minimality (PM), McCarthy and Prince's (1993:117) stem domain of alternative allomorphs (DomAlt), and the affix subcategorization (AffSub) with the ranking of  $\{\text{AffSub} \gg \text{PM}\}$  and DomAlt along with Stump's (2016:77) stem function can explain some of the data, as in Tableau 1. PM, which prohibits words *smaller* than one binary moraic foot, explains the pairing of  $k + uru$  'come + nonpast' over  $*k + u$  (the upper part of the tableau). The subcategorization constraint states that the shorter stem alternant is selected by the nonpast affix. The domain constraint states that the prosodic stem domain of the alternative is the complement of that of the default. In effect, the alternative  $/-uru/$  parses the stem prosodically if and only if the default  $/(r)u/$  cannot do so, which explains the pairing of  $kak + u$  'write + nonpast' over  $*kak + uru$  (the middle part). In contrast, the nonpast forms of the (X)e/(X) stem verbs, as in the paradigm  $\langle \text{tabe}_{adverbial}, \text{tab} + \text{uru}_{nonpast} / * \text{tab} + u, \text{tabe} + \text{ta}_{past}, \text{tabe} + \text{N} / \text{tabe} + \text{raN}_{negative} \rangle$ , cannot be explained even by adding such an existing constraint among the stems or affix' of the forms of paradigms as Uniform Exponence (UE) as in the

Tableau 1: Predictions

		AffSub	PM	UE	DomAlt
/ $\{\text{C}, \text{CV}\} + \{\text{u}, \text{ru}, \text{uru}\} / \text{nonpast}$					
☞	a. Curu b. Cu c. CVru	*!	*!	*	
/ $\dots\text{C} + \{\text{u}, \text{ru}, \text{uru}\} / \text{nonpast}$					
☞	a. ...Cu b. ...Curu				*!
/ $\{\text{(X)e}, \text{(X)}\} + \{\text{u}, \text{ru}, \text{uru}\} / \text{nonpast}$					
☞	a. (X)u b. (X)uru c. (X)eru	*!		*	*!



ranking of {PM  $\gg$  UE} in Tableau 1. The candidate /taburu/ violates the constraint DomAlt whichever the default or basic stem is assumed to be between /tab/ and /tabe/. For example, if /tabe/ is the default stem or the uniform exponence of the stem of the lexeme, the candidates /tabu/ and /taburu/ equally violate UE, as in the lower part of the tableau. The candidate /taburu/ violates DomAlt because the affix alternative is not motivated for the nonpast form of /tab(e)/ because the candidate /tabu/ is not less optimal than /taburu/ except for the computation of the constraint DomAlt. This problem motivates leveling among paradigmatic patterns (or inflectional classes) (Garrett, 2008). Zodak and Bat-El's (2015) theory of leveling among inflectional classes of the Hebrew verb system incorrectly predicts the affix allomorphy of the Japanese dialect. This is because the directionality of leveling is determined by the numbers of the members (the lexical frequencies) of the inflectional classes in their theory. The number of the C-final stem verbs is 56% of the total in the Tokyo dialect in Japanese textbooks, that of V(e)-final stem verbs is 19%, that of the *s/si* stem verbs is 19%, that of the V(i)-final stem verbs is 0.04%, and that of the *k/ko* stem verbs is 0.02%.<sup>1</sup> According to Zodak and Bat-El's (2015) directionality analysis, if the paradigmatic patterns of the (X)e/(X) stem and V-final stem verbs are similar, as researchers will claim below, the directionality of leveling should be from the (X)e/(X) stem verbs to the V-final stem verbs. If this were the case, the paradigmatic pattern of the V-final stem verbs would become as complicated as that of the (X)e/(X) stem verbs, as derived in (2c) from (2a) and (2b). In fact, this is not the case.

- (2) a.  $PP_V$ :  $\langle (X)V, (X)Vru, (X)Vsasuru, (X)Vta, XVraN/(X)VN \rangle$ , e.g.,  $\langle oki, oki + ru, oki + ta, oki + N/oki + raN \rangle$   
 b.  $PP_{Xe/X}$ :  $\langle (X)V, (X)uru, (X)Vsasuru, (X)Vta, (X)VraN/(X)VN \rangle$   
 c.  $PP_{V-by-Xe/X}$ :  $\ast \langle (X)V, \ast(X)uru, (X)Vsasuru, (X)Vta, XVraN/(X)VN \rangle$ , e.g.,  $\ast \langle oki, \ast ok + uru, oki + ta, oki + N/oki + raN \rangle$

### 3 Proposal

The author's proposal of directionality and similarity scale is detailed below.

- (3) a. The directionality of leveling is from a paradigmatic pattern without stem alternation to one with stem alternations.  
 b. 'The more similar the inflectional classes [or paradigmatic patterns] are, the more likely they are to interact in inter-paradigm leveling' [brackets are mine]. (Zodak and Bat-El, 2015: 275)  
 c. The ranking for directionality is Stem Alternation constraint (3a)  $\gg$  Similarity constraint (3b)  $\gg$  Zodak and Bat-El's (2015) Lexical Frequency constraint.  
 d. The degree of similarity between the candidate paradigmatic pattern in question,  $PP_Q$ , and a leveling pattern,  $PP_L$ , is indicated by the sum of the differences between each leveled form of the leveled pattern by the leveling one  $PP_{Q-by-L}$  and its corresponding candidate form of the candidate pattern  $PP_Q$ .

The constraint (3a) confirms Albright's (2005) finding that a pattern of non-alternation is extended in leveling. For the computation of the degree of similarity between two paradigmatic patterns, the four-place analogy is applied to Albright's (2005) derivation of forms from base forms in the clause (3b): 1) identify a derivational rule that derives a form of a morpho-syntactic property from the base form of the leveling paradigmatic pattern and 2) apply the rule to the

<sup>1</sup>The author assumes that the lexical frequencies of the stem type verbs in Japanese textbooks may not be vastly different from the type and token frequencies of the verbs of the four classes in daily conversations.

base form of the paradigmatic pattern in question to derive the counterpart form. It does not matter which verb forms are assumed to be the base forms, as will be discussed later. The difference between an actual and its leveled form will be two if a vowel is existent in one and not in the other, one if the qualities of the vowels are different, and one if a consonant is existent in one and not in the other. Predictions of Which Paradigmatic Pattern Levels Another: By Stem Alternation constraint (3a), the leveling paradigmatic pattern of that of the (X)e/(X) stem verbs is either that of the V-final stem verbs or the C-final stem verbs because only the V-final and C-final stem verbs have no stem alternation among the four types, as can be seen in stem patterns given for each in (1). Employing Zodak and Bat-El's (2015: 275) similarity analysis, the paradigmatic pattern  $PP_L$  to level a pattern in question  $PP_Q$  is the one in which the sum of the differences between (each leveled form of) the leveled pattern by the leveling one  $PP_{Q-by-L}$  and (its corresponding candidate form of) the candidate pattern  $PP_Q$  is the least. The paradigmatic pattern  $PP_V$  (2a) is preferred to pattern  $PP_C$   $\langle XCi, XCu, XCasuru, XC(i)ta, XCaN \rangle$  for the leveling pattern of  $PP_{Xe/X}$  (2b).  $PP_V$ , of which the nonpast forms are not considered, is more similar to or precisely the same as  $PP_{Xe/X}$  than  $PP_C$  is. As the form patterns of  $PP_{Xe/X}$  and the corresponding form patterns of  $PP_V$  are the same, the sum of the differences of the forms in  $PP_{Xe/X-by-V}$  from their counterparts in  $PP_{Xe/X}$  is computed as 0/0. Conversely, the sum of the differences of the forms of  $PP_{Xe/X-by-C}$  (4a) from the corresponding forms of  $PP_{Xe/X}$  is computed as 8/6, as below and the differences computed in (4b).

- (4) a. If  $PP_C$  leveled  $PP_{Xe/X}$ , the leveled paradigmatic pattern  $_{Xe/X-by-C}$  would be  $\langle \underline{(X)V}$ ,  $-$ ,  $\underline{(X)asuru}$ ,  $\underline{(X)(V)ta}$ ,  $\underline{(X)aN} \rangle$ .
- b. The differences of  $PP_{Xe/X-by-C}$  from  $PP_{Xe/X}$  are  $\langle \underline{(X)V} - (X)V$ ,  $-$ ,  $\underline{(X)asuru} - (X)Vsasuru$ ,  $\underline{X(V)ta} - (X)Vta$ ,  $\underline{(X)aN} - (X)VraN/(X)eN \rangle$ , i.e.,  $\langle 0, -, 3, 2, 3/1 \rangle$

The leveled causative form of the (X)e/(X) stem verbs, for example, is *Xasuru* using a four (4)-place analogy for the C-final stem verbs and the (X)e/(X) stem verbs, for Z, (X)V : (X)*asuru* = (X)V : Z because the causative form is derived by concatenating /asuru/ at the end of the adverbial form with the last vowel absent. The difference of the leveled form from the actual form, for example, is (X)*Vsasuru* minus (X)*asuru*, or Vs, or a vowel and a consonant, or 2 + 1 = 3. The sum of the differences of the leveled forms from the actual counterparts is 3 + 2 + 3/1 = 8/6. If the base forms are assumed to be causative forms, for example, the prediction holds true that  $PP_V$  is more similar than or precisely the same as  $PP_{Xe/X}$  than  $PP_C$  is. The leveled paradigmatic pattern  $_{Xe/X-by-C}$  will be  $\langle \underline{(X)Vsi}$ ,  $-$ ,  $\underline{(X)Vsasuru}$ ,  $\underline{(X)Vs(i)ta}$ ,  $\underline{(X)VsaN} \rangle$ . Even with this assumption, because the form patterns of  $PP_{Xe/X}$  and corresponding form patterns of  $PP_V$  are the same, the sum of the differences of the forms in  $PP_{Xe/X-by-V}$  from their counterparts in  $PP_{Xe/X}$  is computed as 0/0. The differences of the forms of  $PP_{Xe/X-by-C}$  from those of  $PP_{Xe/X}$  are  $\langle \underline{(X)Vsi} - (X)V$ ,  $-$ ,  $\underline{(X)Vsasuru} - (X)Vsasuru$ ,  $\underline{XVs(i)ta} - (X)Vta$ ,  $\underline{(X)VsaN} - (X)VraN/(X)eN \rangle$ ,  $\langle 0, -, 3, 3, 3/0 \rangle$ ; the sum is 9/6. Predictions of What is the Leveled Form: Because  $PP_V$  levels  $PP_{Xe/X}$ , the leveled nonpast form of the (X)e/(X) stem verbs is computed as Z by the analogy (X)V : (X)Vru = (X)V : Z, or (X)Vru. The leveled paradigm is  $\langle (X)V$ ,  $\underline{(X)Vru}$ ,  $\underline{(X)Vsasuru}$ ,  $\underline{(X)Vta}$ ,  $\underline{(X)VraN/(X)VN} \rangle$ , for example,  $\langle \text{tabe}$ ,  $\underline{\text{taberu}}$ ,  $\text{tabesasuru}$ ,  $\text{tabeta}$ ,  $\underline{\text{taberaN/tabaN}} \rangle$ . If vowel quality is abstracted or the leveling is only of prosodic structure, the prosodic-structure (PS)-leveled paradigm pattern will be  $\langle (X)V_1$ ,  $\underline{(X)V_2ru}$ ,  $\underline{(X)V_1sasuru}$ ,  $\underline{(X)V_1ta}$ ,  $\underline{(X)V_1raN/(X)eN} \rangle$ . By leveling between paradigmatic patterns, an abstract paradigmatic pattern is created to subsume the previous two similar but parallel and independent paradigmatic patterns. Leveling Constraint and Predictions: More than one form pattern may be morphologically well-formed in some cell of a paradigmatic pattern with stem alternation. The constraint (5a) (PS-L) excludes this type of paradigmatic pattern with stem alternation if it is not prosodically leveled by another paradigmatic pattern, specifically one without stem alternation.

- (5) a. Leveling in Prosodic Structure (PS-L): Assign one violation mark to a paradigmatic pattern  $PP_Q$  if one form pattern of the paradigmatic pattern  $PP_Q$  differs from the counterpart of its leveled paradigmatic pattern by another paradigmatic pattern  $PP_L$ ,  $PP_{Q-by-L}$ .
- b. {AffSub  $\gg$  PM  $\gg$  PS-L  $\gg$  UE}, DomAlt

Tableau 2: Predictions

		AffSub	PM	PS-L	UE	DomAlt
/(X)(e) + {u, ru, uru}/ <i>nonpast</i>						
☞	a. (X)uru				*	
	b. (X)u			*!	*	
	c. (X)eru	*!				
/{C, CV} + {u, ru, uru}/ <i>nonpast</i>						
☞	a. Curu			*		
	b. Cu		*!			
	c. CVru	*!		*	*	

If constraint PS-L is added to the plausible existing constraints with the ranking of the constraint between PM and UE as in (5b), they will make correct predictions for all affixal phenomena, for example, the nonpast forms of the (X)e/(X) stem verbs and those of the C/CV stem verbs, as in the upper and lower parts of Tableau 2. The candidate nonpast form (X)eru violates the subcategorization of the nonpast affix (AffSub) because the nonpast affix selects shorter stem alternants, or (X), but not (X)e. The candidate (X)u does not follow the pattern (X)V<sub>2</sub>ru, thereby violating the PS-L constraint, whereas the candidate (X)uru follows the pattern. Therefore, the candidate (X)uru is optimal. The

paradigmatic pattern with the nonpast form pattern, <(X)V<sub>1</sub>, (X)V<sub>2</sub>ru, (X)V<sub>1</sub>sasuru, (X)V<sub>1</sub>ta, (X)V<sub>1</sub>raN/(X)eN>, is optimal. For example, the paradigm <tabe, taburu, tabesasuru, tabeta, tabeN/taberaN> is optimal. The leveling constraint PS-L is not crucial for the prediction of the optimal nonpast forms for the C/CV stem verb. In OT, all constraints are violable, and for any constraint, the higher it is ranked, the larger is its effect. Because the minimality constraint PM outranks the leveling constraint PS-L, the effect of the minimality constraint, *Curu* and \**Cu*, has a higher priority than that of the leveling constraint, *Cu* and \**Curu* does. Thus, the constraints and ranking with the leveling constraint make the same prediction as those without the leveling constraint in the case of the C/CV stem verbs, such as in the upper part of the tableau.

## References

- Albright, Adam. (2005). The morphological basis of paradigm leveling. In L. J. Downing, T. A. Hall, and R. Raffelsiefen (Eds.), *Paradigms in phonological theory*, 17-43. Oxford: Oxford University Press.
- Garrett, Andrew. (2008). Paradigmatic uniformity and markedness. In Jeff Good (ed.), *Explaining linguistic universals: Historical convergence and universal grammar*, 125-143. Oxford: Oxford University Press.
- Koga, Hiroki. (2023). Another complex phenomenon for Harmonic Serialism and Against Standard Parallel OT. *The Joint Journal of the National Universities in Kyushu. Education and Humanities*, 9(2), No.2
- McCarthy, John J. and Alan Prince. (1993). Prosodic morphology I: Constraint interaction and satisfaction. *Linguistic Department Faculty Publication Series 14* (2001 version).
- Stump, Gregory T. (2016). *Inflectional paradigms: Context and form at the syntax-morphology interface*. Cambridge University Press.
- Zadok, Gila and Outi Bat-El. (2015) Inter-paradigm leveling in Hebrew verbal system, *Morphology*, volume 25 (3): 271–297.

---

# Obscuring morphomic patterns: some evidence from Catalan verbal inflection

Manuel Badal

Universitat de València

---

## 1 Introduction

During the Middle Ages, a new verb class was formed within Catalan 2<sup>nd</sup> conjugation, characterized by the presence of a velar augment (/g/ or /sk/, depending on the verb) at certain cells of the paradigm (about this diachronic process, see, e.g., Pérez Saldanya 1998; Wheeler 2011). The distribution of the velar can be considered “morphomic” (in Aronoff’s 1994 terms), since the systematic appearance of this augment “cannot be aligned with any conceivable coherent semantic, syntactic, or phonological generalization” (Esher and O’Neill 2022: 351). The resulting layout from the analogical velarization process favored an implicative organization of the allomorphs, which made easier the acquisition of the paradigm.

As illustrated in table 1 with the verbs *beure* ‘to drink’ and *créixer* ‘to grow’, the velar augment usually appears in the L-pattern (Maiden 2018: 84), formed by the first-person present indicative and the whole present subjunctive, and PYTA (Menéndez Pidal 1904), formed by the preterite, the old conditional and the imperfect subjunctive, and also in the past participle if this form adopts the regular ending *-ut/-uda* (except for the verb *vendre* ‘to sell’, which presents the non-velarized form *venut* ‘sold’). Root-stressed past participles, such as in the verbs *dir* ‘to say’ or *prendre* ‘to take’, do not adopt the velar: *dit* ‘said’, *pres* ‘taken’.

		<i>beure</i> ‘to drink’ /g/	<i>créixer</i> ‘to grow’ /sk/
L-pattern	1sg present indicative	<i>bec</i> [ˈbek]	<i>cresc</i> [ˈkresk]
	Present subjunctive	3sg <i>bega</i> [ˈbeya]	3sg <i>cresca</i> [ˈkreska]
PYTA	Preterite	3sg <i>begué</i> [beˈɣe]	3sg <i>cresqué</i> [kresˈke]
	Conditional	3sg <i>beguera</i> [beˈɣera]	3sg <i>cresquera</i> [kresˈkera]
	Imperfect subjunctive	3sg <i>begués</i> [beˈɣes]	3sg <i>cresqués</i> [kresˈkes]
Past participle	<i>begut</i> [beˈɣut]	<i>crescut</i> [kresˈkut]	

Table 1. L-pattern, PYTA, and past participle of *beure* ‘to drink’ (exponent of /g/ model) and *créixer* ‘to grow’ (exponent of /sk/ model) in Old Catalan. (In the tenses in which all the forms are velarized, we only offer the 3sg.)

The distribution of the velar augment shown in table 1 is the result of a diachronic change, which led to the “coalescence” (in Maiden’s 2018: 292 terms) of the L and PYTA morphomic patterns. Even with all that, there are certain verbs that did not fully adopt this velar distribution pattern. In this paper, we analyze the causes that can explain the apparently uncoherent spread of velarization in three Catalan 2<sup>nd</sup> conjugation verbs with a remarkable high frequency: *haver* ‘to have’, *ser* ‘to be’ and *voler* ‘to want’. Based on the results obtained from a corpus of texts ranging from the 13<sup>th</sup> to the 19<sup>th</sup> century, we test whether a “coherence” effect took place in the diachronic change of these verbs, as follows from Maiden’s (2018: 3) prediction, “morphological innovations of any kind (e.g. analogical levelling of the alternation, analogical extension of the alternation, creation of novel alternants, introduction of suppletive forms into paradigms) that affect any one of the

paradigmatic cells implicated in the alternation pattern equally and always affect all the others.” Our initial assumption is, thus, that if the two morphemes are psychologically real for speakers (Maiden 2018: 13), all the forms associated to them must undergo the same analogical change. We will show that the corpus data contradict this hypothesis, though cast some light on the factors that can hinder the processes of morphological levelling and regularization.

## 2 Data and methodology

To extract the verb forms, we have set up a corpus comprising Catalan works ranging from the 13<sup>th</sup> to the 19<sup>th</sup> century. We have chosen this period to be able to study the velarized forms from the first Catalan texts until the promulgation of Fabra’s norm; the inclusion in the study of the 20<sup>th</sup> century would have introduced a clearly independent variable of the morpheme influence: the norm impact, with possible effects especially on the analyzed forms that have not been admitted in the standard language, such as *vullc* [ˈvuʎk] ‘I want’. In addition, all the corpus texts belong to the second half of each century, since, methodologically, the objective is to determine the evolution of verb forms at the end of each one.

Once the counts have been made, to elucidate whether the distribution of the non-velarized and velarized forms (variable ‘velarized’) in relation to the centuries (variable ‘century’) is random or not, we have carried out several chi-square tests with SPSS Statistics (IBM Corp. 2019). The chi-square test is based on the comparison of the bivariate frequencies obtained from the data (empirical frequencies) with the frequencies that would result if there were no association between the variables ‘velarized’ and ‘century’ (theoretical frequencies). The test produces two indicators: the  $\chi^2$  value for a two variables distribution –non-velarized and velarized forms– in the periods in which the seven centuries are grouped and the asymptotic significance ( $p$ ). The  $p$ -value is evaluated from the threshold of 0.05, as is usual in the social sciences: when the  $p$ -value is less than 0.05, the probability that the elements have been randomly distributed according to the global frequency between the different groups is lower (less than 5%); in this case, we should discard the null hypothesis according to which the variable ‘century’ does not influence the distribution of verb forms and assume that, on the contrary, the distribution of these forms in the different centuries varies significantly. On the other hand, if  $p$ -value is greater than 0.05, in other words, if the probability of obtaining the real random distribution is greater than 5%, we will accept the null hypothesis and assume that the verb forms have been distributed with the same criterion in the different centuries. Once it has been checked whether the forms are randomly organized or not, it is necessary to test the association between the variables. This data is determined by Cramér’s; this parameter range is 0 to 1: a weak association is 0.1 to 0.2; a moderate one, 0.2 to 0.4; a relatively strong one, 0.4 to 0.6; a strong one, 0.6 to 0.8, and a very strong one, 0.8 to 1 (Rea & Parker 2014: 219).

## 3 Results and discussion

### 3.1 *Haver* ‘to have’

In the verb *haver* ‘to have’, only some velarized forms in the present subjunctive are attested, like *haga* [ˈaɣa] ‘s/he have’, although never became statistically significant ( $\chi^2_{(3)} = 4.990$ ,  $p = 0.172$ , Cramér’s  $V = 0.107$ ). The high frequency of this verb, used mainly as a perfect auxiliary (see IEC 2016), probably favored the preservation of irregular forms over the

centuries (Anshen & Aronoff 1988; Booij 1997: 43; Bybee & Brewer 1980: 218; Rainer 1988). Additionally, since the first person of the present indicative did not adopt the velar consonant, it could not exert the same force as in verbs such as *caure* ‘to fall’ or *deure* ‘to owe’, in which /g/ was first introduced into the first-person present indicative and then extended to the present subjunctive because of the class-stability principle (see Badal 2022). Regarding PYTA, the only tense where changes are observed is the old conditional ( $\chi^2_{(6)} = 112.586$ ,  $p < 0.001$ , Cramér’s  $V = 0.784$ ), with some non-velarized forms in the 19<sup>th</sup> century documented in Valencian, like *havera* [a'vera] ‘s/he would have’. In this variety, the original forms of the imperfect subjunctive, such as *hagués* [a'yes] ‘s/he had’ or *haguesses* [a'yeses] ‘you had’, were no longer used, and the preterite had lost much vitality. Consequently, the PYTA morpheme was blurred, since the only used tense at this time was the old conditional, used then as an imperfect subjunctive (Ridruejo 1985). However, this analogical change is not consolidated at the end of the studied period, and presents dialectal variation, since the forms with the velar consonant (e.g., *haguera* [a'ɣera] ‘s/he would have’) are widely attested at the beginning of 20<sup>th</sup> century (see Alcover & Moll 1929-1932).

### 3.2 *Ser* ‘to be’

As for *ser* ‘to be’, the first person of the present indicative ( $\chi^2_{(6)} = 244.435$ ,  $p < 0.001$ , Cramér’s  $V = 0.910$ ) and the present subjunctive ( $\chi^2_{(6)} = 3323.054$ ,  $p < 0.001$ , Cramér’s  $V = 0.959$ ) present a very similar chronology, since both consolidate the velarization in the 19<sup>th</sup> century. The velarization of the first person of the present indicative (*soc* [ˈsok] ‘I am’) is justified by convergence (Maiden 2005: 139-140) with most 2<sup>nd</sup> conjugation verbs, which adopted /g/ in forms with this tense-mood (e.g., *bec* [ˈbek] ‘I drink’, *dec* [ˈdek] ‘I owe’). In the present subjunctive, the analogical velarization could be due to two factors: the influence exerted by the first person of the present indicative and the tendency to repair the hiatus of the etymological forms (e.g., *si.a* [ˈsia] >> *si.ga* [ˈsiɣa] ‘s/he be’) giving the syllable a simpler structure (Hualde 1992: 383). Regarding PYTA, the third person of the preterite form is the only one coming from Latin *perfectum* in which velarized cases are documented ( $\chi^2_{(6)} = 1251.345$ ,  $p < 0.001$ , Cramér’s  $V = 0.655$ ): *fonc* [ˈfoŋk] ‘s/he was’. It is difficult to come out with a convincing explanation for the evolution undergone by this form: until the 15<sup>th</sup> century, the non-velarized forms (i.e., *fo* ‘s/he was’) are the predominant ones. However, *fonc* equals the frequency of the non-velarized forms in the 16<sup>th</sup> century and reaches its peak in the 17<sup>th</sup>. But in the 18<sup>th</sup> century the trend is reversed again and the forms without the /g/ become again the most common ones. Finally, in the 19<sup>th</sup> century the velarized forms become the minority. The final recession of velarized forms might be attributed to the need for coherence: since over the centuries no other PYTA form adopts the /g/, the velarization process is blocked, since a single PYTA cell with a velar augment is unusual.

### 3.3 *Voler* ‘to want’

In the verb *voler* ‘to want’, the L-pattern resulting from regular sound change (i.e., *vull* [ˈvuɫ] ‘I want’ ~ *vulla* [ˈvuɫa] ‘s/he want’) undergoes an alteration driven by analogical velarization. While Valencian still preserves a homogeneous pattern, the morpheme is obscured in general Catalan. The first person of the present indicative adopts the velar in Valencian only ( $\chi^2_{(4)} = 83.782$ ,  $p < 0.001$ , Cramér’s  $V = 0.794$ ): *vullc* [ˈvuɫk] ‘I want’. However, the analogical velarization of the present subjunctive is widespread in Catalan ( $\chi^2_{(6)} = 356.474$ ,  $p < 0.001$ , Cramér’s  $V = 0.983$ ): e.g., *vulga* [ˈvulya] /*vullga* [ˈvuɫɣa] ‘s/he want’. In the evolution of *voler* in Valencian, the principles of uniformity (Mayerthaler 1987)

and class-stability (Wurzel 1987) take precedence over phonological congruence, given that the L-pattern uniformity is preserved: all forms share the same root-final palatal and present the velar consonant (i.e., *vullc* ~ *vullga*), despite this unusual combination (about this phonological restriction, see Wheeler 1984: §3.4; Hualde 1992: 381; IEC 2016: 25). Even so, in general Catalan the phonological congruence hinders the morpheme homogenization, leading to the L-pattern obscuring, and the generation of a new irregularity: while the first person of the present indicative retains the etymological form, the present subjunctive forms adopt a new allomorph to which the velar is added (i.e., *vull* ~ *vulga* / *vulgui*).

## 4 Conclusions

To sum up, although in most cases the evolution of morphomic patterns is usually coherent because all morpheme cells undergo the same change(s), the present investigation brings to light cases in which this tendency is obscured. More specifically, the results of the study corroborate what some authors have already observed for other phenomena: the more a verb is used, the less susceptible it is to morphological regularization patterns. The data under analysis in this paper, thus, cast light on the fact that sometimes frequency can exert a greater pressure than morphomic coherence in verbal diachronic evolution.

## References

- Alcover, Antoni M. & Francesc de B. Moll. 1929-1932. La flexió verbal en els dialectes catalans. *Anuari de l'Oficina Romànica de Lingüística i Literatura* 2: 79-184; 3: 73-168; 4: 9-104; 5: 9-72.
- Anshen, Frank & Mark Aronoff. 1988. Producing morphologically complex words. *Linguistics* 26. 641–655.
- Aronoff, Mark. 1994. *Morphology by itself: stems and inflectional classes*. Cambridge, MA: MIT Press.
- Badal, Manuel. 2022. La tendència cap a la uniformitat: la velarització dels radicals palatals dels verbs de la segona conjugació del català. *eHumanista/IVITRA* 21. 469–481.
- Booij, Geert. 1997. Autonomous morphology and paradigmatic relations. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1996*, 35–53. Dordrecht: Springer.
- Bybee, Joan & Mary Alexandra Brewer. 1980. Explanation in Morphophonemics: Changes in Provençal and Spanish Preterite Forms. *Lingua* 52(3–4). 201–242.
- Esher, Louise & Paul O'Neill. 2022. The Autonomy of Morphology. In Adam Ledgeway & Martin Maiden (eds.), *The Cambridge Handbook of Romance Linguistics*, 346–370. Cambridge: Cambridge University Press.
- Hualde, José Ignacio. 1992. *Catalan*. New York: Routledge.
- IBM Corp. 2019. *IBM SPSS Statistics for Windows (26.0)*. Armonk, NY: IBM Corp.
- IEC = Institut d'Estudis Catalans. 2016. *Gramàtica de la llengua catalana*. Barcelona: Institut d'Estudis Catalans.
- Maiden, Martin. 2005. Morphological autonomy and diachrony. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2004*, 137–175. Dordrecht: Springer.
- Maiden, Martin. 2018. *The Romance Verb: Morphomic Structure and Diachrony*. Oxford: Oxford University Press.
- Mayerthaler, Willi. 1987. System-independent morphological naturalness. In Wolfgang U. Dressler, Willi Mayerthaler, Oswald Panagl & Wolfgang U. Wurzel (eds.), *Leitmotifs in Natural Morphology*, 25–58. Amsterdam: John Benjamins Publishing Company.

- Menéndez Pidal, Ramón. 1904. *Manual elemental de gramática histórica española*. Madrid: Librería General de Victoriano Suárez.
- Pérez Saldanya, Manuel. 1998. *Del llatí al català: morfosintaxi verbal històrica*. València: Universitat de València.
- Rainer, Franz. 1988. Towards a theory of blocking: the case of Italian and German quality nouns. In Geert Booij & Jaap van Marle (eds.), *Vol 1 1988*, 155–186. Berlin & Boston: De Gruyter Mouton.
- Rea, Louis M. & Richard A. Parker. 2014. *Designing and Conducting Survey Research: A Comprehensive Guide*. San Francisco: Jossey-Bass.
- Ridruejo, Emilio. 1985. La forma verbal en *-ra* en valenciano. In *Linguistique comparée et typologie des langues romanes. Actes du XVIIème Congrès International de Linguistique et Philologie Romanes (Aix-en-Provence 1983)*. Vol. II, 437–448. Aix-en-Provence: Université de Provence.
- Wheeler, Max W. 1984. La conjugació valenciana: geografia, diacronia i psicologia. In *Miscel·lània Sanchis Guarner, I. Estudis en memòria del professor Manuel Sanchis Guarner: estudis de llengua i literatura catalanes*, 409–419. València: Universitat de València.
- Wheeler, Max. W. 2011. The Evolution of a Morpheme in Catalan Verb Inflection. In Martin Maiden, John Charles Smith, Maria Goldbach, & Marc-Olivier Hinzelin (eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, 182–209. Oxford: Oxford University Press.
- Wurzel, Wolfgang U. 1987. System-dependent morphological naturalness in inflection. In Wolfgang U. Dressler, Willi Mayerthaler, Oswald Panagl & Wolfgang U. Wurzel (eds.), *Leitmotifs in Natural Morphology*, 59–96. Amsterdam: John Benjamins Publishing Company.



---

# From competing patterns to competing structures: Verbal constructions based on loanwords in Hebrew

Lior Laks

Bar-Ilan University

---

## 1 Introduction

The study examines the competition between morphological and periphrastic structures in the expression of verbal meaning, based on loanwords in Hebrew. As demonstrated in the online examples below, the meaning of the verb ‘talkback’ is expressed morphologically using the infinitive form of the verb *tikbek*, which is formed in the *CiCeC* pattern (1a). In contrast, the same meaning is expressed periphrastically (1b) by using the infinitive form of the verb *katav* ‘write’ with the loanword ‘talkback’. The same meaning can be expressed either by a single word using a morphological process, or by a multi-lexemic expression, and the two structures compete for the same meaning.

- (1) a. kol exad yaxol **letakbek**  
‘Everybody can write a talkback’

<https://www.dwh.co.il/226-dwhcoil/1411-%D7%A9%D7%93%D7%A8%D7%95%D7%92-%D7%94%D7%90%D7%AA%D7%A8-%D7%A9%D7%9C%D7%91-1-1>

- b. kol adam yaxol **lixtov tokbek**  
‘Every person can write a talkback’

<http://www.oritkamir.org/%D7%97%D7%95%D7%A4%D7%A9-%D7%94%D7%91%D7%99%D7%98%D7%95%D7%99-%D7%99%D7%95%D7%AA%D7%A8-%D7%9E%D7%93%D7%99-%D7%91%D7%99%D7%98%D7%95%D7%99-%D7%A4%D7%97%D7%95%D7%AA-%D7%9E%D7%93%D7%99-%D7%97%D7%95/>

Semitic morphology relies highly on non-concatenative morphology, namely the combination of root and pattern. The patterns indicate the prosodic structure of verbs, their vocalic patterns and their affixes (if any). For example, the verb *siper* ‘tell’ is formed in *CiCeC*, and *hitraxec* ‘wash oneself’ in *hitCaCeC*. The phonological shape of a verb is essential for determining the shape of other forms in the inflectional paradigm (Berman 1978; Schwarzwald 1981, Bolozky 1978, Ravid 1990, Bat-El 1994, Aronoff 1994). Most studies of Hebrew verb formation focus on verbal patterns, the relations between them and competition between them namely, the criteria for selection of one pattern and not another. For example, transitive verbs are typically formed in *CiCeC*, e.g. *tilfen* ‘telephone’, while intransitive verbs like reflexives and inchoatives are formed in *hitCaCeC*, e.g. *hitmagnet* ‘become magnetized’. The current study examines a different type of competition, namely the competition between using one of the verbal patterns or a periphrastic construction. I will show that the selection between the two structures can be partially predicted based on the interaction between morpho-phonological and lexical-semantic criteria. The study is based on online searches and the HebTenTen corpus.

## 2 Morpho-phonological criteria

### 2.1 Number of syllables

Most studies have focused on the competition between patterns. Examine the verbs in (2) which are derived from the loanwords *debug* and *spam*.

- (2) a. *dibag* - *dibeg* / \**hidbig* 'debug'  
 b. *spam* - *hispim* / \**sipem* 'send a spam'

The verb *dibeg* is formed in CiCeC and not *hiCCiC* (\**hidbig*), while the verb *hispim* is formed in *hiCCiC* and not *CiCeC* (\**sipem*). The selection of *hiCCiC* allows preserving the consonant cluster of *spam*, and therefore, such formation is more faithful to the base. *dibeg* is derived from a base with no consonant cluster, and a formation of *hiCCiC* would result in an undesired cluster. This is an example of faithfulness constraint that determines the competition between patterns, and as a result of such faithfulness, the structural relations between the base and derived verb is more transparent. Various studies have demonstrated the importance of preserving properties of the base in Hebrew verb formation (Boložky 1978, Bat-El 1994, 2017, Ussishkin 2005, Faust 2015, among others). The current study takes this matter one step further, arguing that low structural transparency can also block verb formation and bring about preference for periphrastic constructions (see Halevy-Nemirovsky 1998), where the loanword remains intact. This is primarily related to the number of syllables of the base. Most verbs are derived from bases that do not exceed two syllables. In case of 3 or more syllables, at least one vowel has to be deleted, making the derived verb less faithful to the stem. For example, the verb *kitleg* 'put in a catalogue' is derived from *katalog* 'catalogue' and its formation involves deletion of second vowel. Such cases are possible, but are less frequent in comparison with stems with less syllables. In contrast, the formation of verbs like *tikbek* based on *tokbek* 'talkback' (1a) involves only changing one of the vowels and the syllabic structure remains intact. The word *fotošop* 'Photoshop', for example, has no derived verb like \**fiššep*. Instead, the construction *asa fotošop* 'do Photoshop' is used. There seems to be no semantic reason for not deriving such a verb, apart from the low structural transparency. It is important to note that this reflects tendencies rather a dichotomy. Verb formation based on words with more than two syllables is possible, but the fact that most cases of the lack of verb formation is when there are more than two syllables is not a coincidence. Loan words like *babysitter*, *relocation*, *taekwondo*, *filibuster* and *paparazzi* are all common in periphrastic verbal constructions in Hebrew, and have no derived verbs, and the more syllables there are the smaller the chances of verb formation.

## 2.2 Non-native suffixes

Loanwords with typical non-native suffixes do not have derived verbal counterparts. This is mostly found in loanwords with the English suffix *-ing*. Hebrew speakers identify these words as typical loanwords, and as a result they are less likely to be integrated into the morphological system. A loanword like *šoping* 'shopping' does not have a verbal counterpart like \**šipeng*, but only a periphrastic construction like *asa šoping* 'do shopping'. Similarly, loanwords like *mingeling* 'mingling', *gosting* 'ghosting' and *fišing* 'fishing', which are highly frequent in verbal periphrastic constructions, but do not have derived verbs. This suggests that the morphological mechanism is sensitive to the morphological structure of loanwords, and in cases where it identifies typical non-native morphological elements, it tends not to integrate such words into the verbal system.

### 3 Lexical-semantic criteria

#### 3.1 Semantic transparency

Low semantic transparency blocks periphrastic formation. In such cases, the meaning of the derived verb is not transparent in relation to the base. For example, the verb *firmet*, derived from *format* ‘format’ does not mean formatting in general but formatting a computer. The noun *format* is borrowed into Hebrew but in a more general sense, not restricted to the domain of computers. The verb *firmet* has no periphrastic alternative like *šina format* ‘change format’ or *sam be-format* ‘put into a format’ that would match the context of computer formatting. As a result, only the morphological construction is used. Similarly, the verb *tirped* ‘ruin (plans)’ is derived from *torpedo* ‘torpedo’, but has a metaphorical meaning, which cannot be expressed via a periphrastic construction with the word *torpedo*.

#### 3.2 Lexical category

The selection between morphological and periphrastic formation can be partially predicted based on the lexical category of the base. In case the base is a verb, morphological formation is obligatory. Verbs that are borrowed directly into Semitic languages must have a pattern. The verb *hitfayed* ‘fade’, for example, cannot have a periphrastic alternative as the word *fade* itself is not used in Hebrew because it is a verb. Nouns are borrowed directly into Hebrew without morphological adaptation (only phonological), and therefore can be the base for both morphological and periphrastic formation, based on the criteria discussed so far. Adjectives are an intermediate category between nouns and verbs with respect to borrowing (Ravid 1990, Schwarzwald, 2013). Some adjectives are borrowed directly with no morphological adaptation, e.g. *snob* ‘snob’. Most borrowed adjectives undergo morphological adaptation of three types: (i) affixation of *-i*, which is a typical Hebrew adjectival suffix, e.g. *efektiv-i* ‘effective’; (ii) truncation of a final consonant, which results in an *i* ending adjective, e.g. *komi* ‘comic’; and (iii) templatic formation, e.g. *medupras* ‘depressed’, which is formed in the *meCuCaC* pattern. So nouns never undergo morphological adaptation, verbs are systematically integrated into the morphological system of root and pattern, and adjectives are in the middle. This intermediary status of borrowed adjectives is also manifested in the selection between morphological and periphrastic constructions to express a verbal meaning. Most adjectives have periphrastic verbal constructions. For example, the adjective *larj* ‘large (generous)’ is used in the construction *nihya larj* ‘become large’, while there is no verbal counterpart like *\*hitlarej*. Similarly, borrowed adjective like *targi* ‘tragic’ and *senili* ‘senile’ have no verbal counterparts. Since most borrowed adjectives undergo some type of morphological adaptation, they are perceived as derived entries and there is a tendency to avoid further derivations, and therefore verb formation is relatively rare. In contrast, in case the base is a borrowed noun, both constructions can be found. For example, the noun *?obsesya* ‘obsession’ is the base for the formation of the verb *hi?tabses* ‘become obsessed’ and the periphrastic construction *haya be-?obsesya* ‘be in an obsession’.

Many studies examined the competition between morphological and periphrastic structures from different points of view (see for example, Haspelmath 2000, Kiparsky 2005, Booij 2010, Corbett 2013, Bonami 2015, Aronoff 2016, Rainer 2016, Štekauer

2016, Masini 2019, among many others), but few studies have addressed it with respect to Semitic morphology, especially in derivation. The study proposes one step in this direction, shedding light on the criteria for the selection between morphological and periphrastic constructions to express verbal meaning. Morpho-phonological criteria block morphological formation due to low structural transparency between the base and the derived verb and the existence of non-native suffixes, which make morphological adaptation more difficult. Low semantic transparency tends to block periphrastic formation, as there are cases with no alternative periphrastic construction that would express the same meaning of the derived verb. In addition, the lexical category of the base provides partial prediction with respect to the possibility to employ wither construction.

#### 4 References

- Aronoff, Mark. 1994. *Morphology by Itself*. Cambridge: MIT Press.
- Aronoff, Mark. 2016. Competition and the lexicon. In A. Elia, C. Iacobino & M. Voghera, *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della società di linguistica Italiana*. Roma: Bulzoni Editore. 39-52.
- Bat-El, Outi. 1994. Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory* 12. 572-596.
- Bat-El, Outi. 2017. Word-based items-and processes (WoBIP): Evidence from Hebrew morphology. In C. Bower, L. Horn, & R. Zanuttini (eds.), *On Looking into Words (and beyond)*, 115-135. Berlin: Language Sciences Press.
- Berman, Ruth A. 1978. *Modern Hebrew structure*. Tel Aviv: University Publishing Projects.
- Bolozky, Shmuel. 1978. Word formation strategies in Modern Hebrew verb system: denominative Verbs. *Afroasiatic Linguistics* 5, 1-26.
- Booij, Gert E. 2010. *Construction Morphology*. Oxford: Oxford University Press.
- Bonami, Olivier. 2015. Periphrasis as collocation. *Morphology* 25, 63-110.
- Faust, Noam. 2015. A novel, combined approach to Semitic word-formation, *Journal of Semitic Studies* 15
- Halavay Nemirovsky, Rivka. 2008. Complementary distribution of single vs. expanded lexical units in Modern Israeli Hebrew. *Leshonenu* 3/4. 293-309.
- Haspelmath, Martin. 2000. Periphrasis. In G. Booij, C. Lehmann & J. Mugdan (eds.), *Morphology: An International Handbook on Inflection and Word-Formation*. 654-664. Berlin: de Gruyter.
- Kiparsky, Paul. 2005. Blocking and periphrasis in inflectional paradigms. *Yearbook of Morphology* 2004, 113-135.
- Masini, Francesca. 2019. Competition Between Morphological Words and Multiword Expressions. In F. Rainer, F. Gardani, W. Dressler, & H.C. Luschützky (eds.), *Competition in Inflection and Word Formation*. Berlin: Springer
- Rainer, Franz. 2016. Blocking. In M. Aronoff, *The Oxford Research Encyclopedia of Linguistics*.
- Ravid, Dorit. 1990. Internal structure constraints on new-word formation devices in Modern Hebrew. *Folia Linguistica* 24. 289-346.
- Schwarzwald, Ora R. 1981. *Grammar and reality in the Hebrew verb*. Bar Ilan University.
- Schwarzwald, Ora R. 2013. The typology of nonintegrated words in Hebrew. *SKASE Journal of Theoretical Linguistics* 10(1). 41–53.
- Štekauer, Pavel. 2016. Compounding from an onomasiological perspective. In P. ten Hacken (ed.), *The Semantics of Compounding*. Cambridge: Cambridge University Press. 54–68.
- Ussishkin, Adam. 2005. A fixed prosodic theory of nonconcatenative templatic morphology. *Natural Language and Linguistic Theory* 23. 169-218.

---

# Defining a Framework for Semantic Categories for Turkish Nominal Morphemes

*Yağmur Öztürk*

Centre de recherches interdisciplinaires et  
transculturelles,  
Université de Franche-Comté

*Izabella Thomas*

Centre de recherches interdisciplinaires et  
transculturelles,  
Université de Franche-Comté

*Snejana Gadjeva*

Centre de recherches Europes-Eurasie,  
Institut National des Langues et  
Civilisations Orientales

---

## 1 Introduction

As an agglutinative language, the main noun formation process of Turkish is suffixation. The roots, whether simple or complex, undergo almost no alternations during the process apart from a few exceptional ones. (1) is an example of a derived noun, with *göz* (en. eye) being the root to which several derivational morphemes are attached.

(1)	Turkish	English
	<i>göz</i>	eye
	<i>göz-lük</i>	eyeglasses
	<i>göz-lük-çü</i>	optician
	<i>göz-lük-çü-lük</i>	occupation of an optician

Despite this rather regular word formation process, there are very few morphological analysers processing derivational forms developed in the fields of NLP (i.e. Çöltekin, 2014), not to mention morphosemantic analysers.

In the linguistic field, several studies and language teaching books provide a description of derivational morphemes. Nevertheless, none offer a comprehensive and systematic description of these morphemes (Bazin, 1994; Adalı, 2004; Göksel & Kerslake, 2005; Korkmaz, 2014; etc) especially in regard to the description of their semantics. Some works propose a semantic study of the morphemes (Bilgin et. al, 2003; Öztürk, 2016) but as it is not the main focus of those works, they seem insufficient to be used as a base in our research.

Our research aims to systematise the semantic description of nominal morphemes in Turkish, with the goal of creating a computerised derivational resource. This line of research has been conducted in many languages within theoretical linguistics as well as the Natural Language Processing field (Talamo et al., 2016; Bagasheva, 2017; Namer et al., 2019; etc). The lack of such research in Turkish is the primary motivation behind our study.

The main question we are concerned with is how to systematise the semantic description of morphemes. How can we turn the glosses and semantic explanations, unique to each author in the studies mentioned above, into a systemic unified set of categories? A systemic description implies a depiction of, or an attempt to depict, the semantic regularities in an analysis of a derivational lexicon.

This study presents the process of establishing this set of semantic categories. First, we tried to take advantage of existing works in this field, even if none have been proposed specifically for the description of the Turkish language. In comparative semantic research in derivational morphology of Bagasheva (2017), however, Turkish is part of the languages described. As briefly discussed in section 2, this set of universal categories proposed by Bagasheva does not meet our expectations for a systemic description of the semantics of nominal morphemes. That is why we directed our research towards another resource, Wordnet (Miller et al. 1990; Fellbaum, 1998), which was not specifically designed for the description of morphosemantics, but which has been applied in recent research in French morphosemantics (Namer et al., 2019).

## 2 Semantic categories for affixes in Bagasheva (2017)

One of the most recent cross-linguistic research in semantics of derivational morphology is Bagasheva's work presented in Comparative semantic concepts in affixation (2017). She defines 51 semantic categories for both local (in the sense of one language), and comparative research, applied to different types of derivation.

In Bagasheva's paper, the set of semantic categories is presented in an alphabetically ordered table. The table is made of three columns: the name of the concept (Comparative semantic concept), then a short definition (Emergent meaning) of the concept and one or two examples usually in Bulgarian and English (2).

- (2) a. ABILITY | Possibility to be processed in a particular way | Eng. readable **readability** / Bul. **čativen** **četivnost**  
b. AGENT | Performer of an activity/ Name of a profession, job, title or permanent activity | Eng. killer / Bul. **ubiec** 'killer'; **pekar** 'baker'

The proposed semantic categories are of different levels of granularity: from basic and general concepts, like ABILITY and AGENT, see e.g. (2), to more specific concepts with more detailed definitions, e.g. (3). Nevertheless, in concrete terms, the hierarchical differences between the semantic categories are neither formally established by the author, nor visible in their description.

- (3) a. FEMALE | Female representative of a human type/profession | Eng. actress / Bul. **čistnica** 'woman fastidious about cleanliness'  
b. UNDERGOER | Entity that undergoes an action that changes its state | Saami **čuhppojuvvot** 'to be cut (of somebody)'

After a test phase on 50 items, we concluded that this set was unsuitable for the description of Turkish nominal morphemes in our framework. The test phase was divided into 4 steps (but we will not explain in detail due to limited space): 1. selection of the categories that fits or may fit in the nominal derivational process in Bagasheva's set, 2. creation of a corpus of 50 derived Turkish nouns for annotation, 3. annotation of the morphemes in the corpus, 4. verification of the annotation and observation of the results presented in Table 1.

Table 1: Results of the corpus annotation

Match to a semantic category	Number of matches	% of matches
Perfect match	25	50%
Several matches	7	14%
No match	18	36%

## 3 Methodology for the definition of semantic categories

From the observation mentioned in 2., we have been able to define the properties of our set of semantic categories for the description of Turkish nominal morphemes. In the following section, we will first present those. Then we will present our baseline, WordNet adapted to the description of Turkish morphosemantics.

### 3.1 Properties of the set

We have defined properties for the set of semantic categories, 4 in number. The following subsections present and discuss those properties.

### **3.1.1 An open-source project**

Another side of our research is that we aim to be part of open research. This means that the finalised resources will be available for any usage in the fields of linguistics and NLP. In our study case, this research is part of a wider project and will be integrated in a morphosemantic analyser for assisted learning of derived nouns in Turkish.

This aspect of our research is an important feature to take into account in the process of establishment because it has a direct impact in the way we define our semantic categories.

Consequently, it is the main criteria in the choice of its modelisation for the resulting computerised resource we created. An ontology is by definition coherent in its structure and can be shared for different purposes. It is a modelisation of knowledge understandable by a community of people and readable by computerised resources. That is why we decided to implement the set of categories in an ontological structure we called Semantürk which aims to provide a comprehensive and structured database of semantic information for nominal derivatives in the Turkish language.

### **3.1.2 A hierarchically structured and organised set**

A simple alphabetically ordered set is not enough to render the semantics of morphemes. Therefore, the most striking criteria is to have an organised and structured set of semantic categories. It is necessary for the semantic categories to be linked to each other in a hierarchical structure as:

- it has different levels of granularity; the hierarchy helps bring these granularities in the semantics of the morphemes out;
- moreover, it guarantees a match for all of the morphemes under study since there is a possible fallback if a morpheme does not match with a really specific semantic category.

### **3.1.3 Non-ambiguous categories**

The most important feature is minimum ambiguity: the semantic categories have to be interpretable. The aim is to describe the semantics of the morphemes in order to comprehend the sense the morphemes convey in a derivational process. So we paid considerable attention to define the categories without ambiguity.

Another reason is related to the first property we presented in 3.1.1: this research has to be exploitable for any usage. As we previously said, in the case of our usage, the semantic categories are to be integrated in a morphosemantic analyser for Turkish learners. So the categories will be described in two aspects, first with a linguistic identifier and a definition accessible for non-linguist users.

### **3.1.4 Usage in other languages**

To have a set of categories applicable, or at the very least partly applicable, to other languages would be an added value as future works in comparative morphosemantics would be possible. The hierarchical nature of this research would, a priori, render such a work possible since not all languages have the same levels of granularity in their morphosemantics.

## **3.2 Discussion: WordNet for a morphosemantic description?**

Wordnet is a lexical database, applied and adapted to a large variety of languages. The semantic components present semantic primes called Unique Beginners and are 25 in number. They were not designed specifically for the description of the morphosemantics of languages as in Bagasheva (2017) but structured to form pairs of words called synsets that are semantically related. Initially built for English, WordNet has since been applied and adapted to many languages around the world and is well-established research.

A recent work Demonext (Namer et al., 2019) used Wordnet's set of semantic categories to describe French morphosemantics. It seems to offer a sufficient granularity in order to describe the morphosemantics of derived words in French. Observing morphologically related synsets, alternations in the meaning can be rendered by the attribution of Unique Beginners to each element of the pair. (4) is an example from Huguin et. al (2022) who worked on the semantic component of Demonext words. It presents the alternation mentioned for words derived with the morpheme *-eur*, alternating from ARTIFACT to PERSON as in (4).

(4) French	code; codeur {ARTIFACT; PERSON}
English	code; coder {ARTIFACT; PERSON}

Inspired by this work, we decided to test and adapt the WordNet's set for the description of derived nouns in Turkish as it meets the criteria discussed in 3.1. The semantic components of Wordnet are constructed in the hierarchical principle as well, with different levels of granularity. We adapted them to match the description of Turkish nominal morphemes.

## References

- Adalı, Oya. 2004. *Türkiye Türkçesinde Biçim Birimler*. Istanbul: Papatya.
- Bagasheva, Alexandra. 2017. Comparative semantic concepts in affixation. In Salvador Valera Hernández & Juan Santana Lario (eds.), *Competing Patterns in English Affixation*, 33-65. Peter Lang.
- Bazin, Louis. 1994. *Introduction à l'étude pratique de la langue turque*. Paris: Librairie d'Amérique et d'Orient.
- Bilgin, Orhan & Kemal Oflazer. 2004. Morphosemantic Relations In and Across Wordnets. In Petr Sojka, Karel Pala, Christiane Fellbaum & Piek Vossen (eds.), *Proceedings of the Second International WordNet Conference (GWC 2004)*, 72-78. Brno, Czech Republic: Masaryk University.
- Çöltekin, Çağrı. 2014. A set of open source tools for Turkish Natural Language Processing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, *Database. Language, Speech, and Communication*. Cambridge, London: The MIT Press.
- Göksel, Aslı & Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar (Routledge Comprehensive Grammars)*. London: Routledge.
- Huguin, Mathilde, Lucie Barque, Pauline Haas, Fiammetta Namer, Delphine Tribout. 2022. *Guide d'annotation Demonext: Typage lexical des noms du français*. <https://hal.science/hal-03638962>
- Korkmaz, Zeynep. 2014. *Türkiye Türkçesi Grameri Şekil Bilgisi*. Türkiye: Türk Dil Kurumu editions.
- Miller, George. A., Richard Beckwith, Christiane Fellbaum, Derek Gross, & Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3:4. 235-244.
- Namer, Fiammetta, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout & Delphine Tribout. 2019. Demonette2 - Une base de données dérivationnelles du français à grande échelle : premiers résultats. In Emmanuel Morin, Sophie Rosset & Pierre Zweigenbaum (eds.), *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, 233-243. Toulouse, France: ATALA.
- Öztürk, Seda. 2016. *Création et reconnaissance de néologismes par méthode de suffixation*. Master's thesis, Université de Franche-Comté.
- Talamo, Luigi, Chiara Celata & Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure* 9:1. 72-102. doi:10.3366/word.2016.0087.



---

# On the polysemy of derivational exponents

*Bernard Fradin*

Laboratoire de Linguistique Formelle, Université Paris Cité & CNRS

---

## 1 Typical cases

Haspelmath (2002) mentions a well-known fact about French language, namely that suffixing *-ier* to the stem of nouns that denote a fruit allows one to derive the name of the plant which yields this fruit. Examples (1) illustrate this possibility and (2) expresses the meaning it involves in the form of an inference.

- (1) *pomm-ier* ‘apple-tree’, *cocot-ier* ‘coconut palm’, *cassiss-ier* ‘blackcurrant bush’
- (2) **Fruit tree** If N denotes a kind of fruit, then N-ier denotes the plant that produces that fruit e.g. *poire* ‘pear’ / *poirier* ‘pear-tree’.

However many other nouns suffixed by *-ier* exist and have a completely different meaning:

- (3) a. **Producer / trader** *bijout-ier* ‘jeweller’, *céréal-ier* ‘cereal farmer’, *chemis-ier* ‘shirt maker’  
b. **Hunter** *renard-ier* ‘fox-hunter’, *canard-ier* ‘duck hunter’, *loutr-ier* ‘otter hunter’  
c. **Container** *sucr-ier* ‘sugar bowl’, *chèqu-ier* ‘checkbook’, *plum-ier* ‘pencil box’

An inferential account can be devised for these nouns too, as long as the base supplies us with the information needed to specify the meaning associated with these derived lexemes.

- (4) **Producer/trader** If N denotes a kind of artefact, then N-ier denotes the person who produces, sells or uses that artefact e.g. *clou* ‘nail’ / *cloutier* ‘nail maker’
- (5) **Hunter** If N denotes a wild animal species, then N-ier denotes the agent who hunts that species e.g. *renard* ‘fox’ / *renardier* ‘fox-hunter’
- (6) **Container** If N denotes a substance or an object having a specific use, then N-ier denotes the container where that substance or object is kept e.g. *sucré* ‘sugar’ / *sucrier* ‘sugar bowl’

This state of affairs corresponds to the situation where one exponent is connected with multiple meanings (1 form  $\leftrightarrow$  n meanings). This situation seems to support the idea that the affix is polysemic. However, it is far from easy to make explicit the various meanings *-ier* would be associated with. Even though this suffix can be used to form nouns that denote agents, trees, containers, etc. it is impossible to say that *-ier* is intrinsically associated with the contents ‘agent’, ‘tree’, ‘container’ etc. without giving the conditions that trigger these respective readings. This is exactly what the inferential approach sketched above does. But this approach clearly shows that the formulation of the appropriate derived meaning depends on two elements: the meaning of the base and the presence of the derivational exponent in question. The contribution of the latter to the derived meaning seems rather tenuous however, compared with that of the base; so tenuous that it eludes any formulation. What is certain is that the derivational exponents in question are not associated with any fixed actual meaning. As a result, these affixes cannot be morphemes and have to be analyzed as morphs (Crysmann & Bonami,

2015; Haspelmath, 2020). They cannot be polysemous insofar as they cannot and need not be correlated with any identifiable meaning. They reveal a meaning without positively bringing with themselves a part of that meaning.

Many affixes behave like *-ier*: they form lexemes associated with meanings that can be formulated through an inference which takes advantage of the meaning of the base:

- (7) **Place of living** If N denotes a domestic animal, N-erie denotes the place where the animal is raised or kept e.g. *chèvre* ‘goat’ / *chèvrerie* ‘goat farm’.
- (8) **Place of manufacturing** If N denotes an agent who exercises an activity, N-erie denotes this activity or the place where it takes place e.g. *coutelier* ‘cutler’ / *coutellerie* ‘cutlery industry’; ‘cutlery shop, works’.

## 2 On the origin of derived meaning

The meaning of the derived nouns mentioned up to now cannot be obtained through the combination of the meaning of their parts insofar as the derivational exponent lacks any identifiable meaning. I assume that the derived meaning comes from the lexical networks the derived lexemes (words) belong to. If we take seriously the word and paradigm approach, then the inferential approach of derivational meaning is self-evident. Blevins (2016, 170) claims that “Paradigmatic relations (...) operate over larger sets of words (...) It is the affiliation with these larger sets of forms that principally constrains uncertainty in the association between individual word-forms and grammatical properties”. In derivation, the uncertainty is constrained by the fact that a given word, or more appropriately lexeme, belongs to a given morphological (or lexical) series. The inferences expressing the derivational meaning are rooted in morphological derivational series and the items forming a series have distinct syntactic distributions. For instance the word *sucriers* has different meanings in (9a) and (9b), respectively ‘sugar bowl’ and ‘sugar manufacturer’, because it is correlated with a lexeme that belongs either to derivational series (10a) or (10b).

- (9) a. L’analyse des possibilités de l’éthanol (...) fait clairement apparaître que les **sucriers** et les distillateurs ne contribueraient que modestement à ce dessein national. (www.persee.fr > doc > rei 0154-3229 1981)
- b. (...) verres de couleur pour les vitraux d’églises et un verre ressemblant à une porcelaine demi-transparente pour les **sucriers** et les compotiers (www.racinescomtoises.net > histoire de malbouans)
- (10) a. *houblon* ‘hop’ / *houblonnier* ‘hop farmer’, *betterave* ‘beetroot’ / *betteravier* ‘beetgrover’, *céréale* ‘cereal’ / *céréaliier* ‘grain farmer’, *pétrole* ‘oil’ / *pétrolier* ‘oil man’
- b. *cendre* ‘ash’ / *cendrier* ‘ash tray’, *plume* ‘nib’ / *plumier* ‘pencil box’, *légume* ‘vegetable’ / *légumier* ‘vegetable dish’, *chèque* ‘check’ / *chéquier* ‘checkbook’

The meaning is built in discourse through sentences such as (9a) and (9b) and the semantic inferences they involve; it comes from outside the word/lexeme, and affix *-ier* only plays a role of trigger, if any, in this process. Sentences (9) illustrate how the syntactic distribution of the relevant items in each series is different.

Two points emerge: (i) the main semantic source of the derived meaning is the base; (ii) this base is supposed to have a sufficiently rich ontology for derived meanings to be specified without problem. It has been argued that nouns are categories of this type (Vicente, 2018; Millikan, 2000): they denote kinds, which means objects endowed with a rich ontology, and

can aggregate content through their use in discourses. “We can draw lots of inferences based on our kind-concepts because they store lots of information. In contrast, concepts of properties or events are informationally ‘flat’” (Vicente, 2018). Verbs would be typical examples of these ‘informationally flat’ categories. They denote eventualities (events or states) of various types, that is basically relations involving variables the role of which in the relation is usually specified through semantic roles (AGT, PAT, INS, etc.). Events are generally associated with a scale (e.g. change-of-state, motion, etc.), which allows us to describe the aspectual properties of the verbal relation, notably its degree of telicity and affectedness (Krifka, 1998; Beavers, 2011, 2013). It should be noted that these properties can only be identified at the phrase or sentence level (Dowty, 1979; Verkuyl, 1993; Rothstein, 2007), when the constructions that the verb heads are unfolded. However rough, this presentation allows us to address two important issues: can the analysis proposed in §1 be extended to derivations from a verbal base? Can the affixes used in these derivations be polysemous?

### 3 Extension of the account

The semantics of nouns derived from verbs is directly correlated with the variables appearing in the semantic representation of the verb. For instance, *-eur* builds agent denoting nouns that are anchored to the agent variable  $x$  of agentive verbs, whereas a subset of derived nouns in *-oir* is linked with a variable which corresponds to the landmark of a spatial (inessive) relationship, as illustrated in (11) with *nageur* ‘swimmer’ and *lavoir* ‘wash-house’ respectively.

- (11) a.  $nageur' = \lambda x \exists e. [swim'(x, e) \wedge AGT(x)]$   
 b.  $lavoir' = \lambda z \exists x y \exists e. [wash'(x, y, e) \wedge AGT(x) \wedge PAT(y) \wedge LOC(e, in'(z))]$

With these exponents the meaning depends on the role assigned to the variable in the verbal construction. It is impossible to adopt an inferential approach to the meaning of the derived N, as we did for *-ier*, inasmuch as there is no element endowed with a rich ontology in the semantic representation of the base V. Besides, there is no need for that.

Contrary to what we saw with *-ier*, affixes *-eur*, *-oir* do not limit themselves to reveal the derived noun’s meaning taking advantage of the content of a base N. They have a proper meaning given by the property captured by the lambda formula headed by the variable that has been selected; it reads ‘X such as she has the property to swim’ for *nageur*, and ‘Z which is a place where X washes Y’ for *lavoir*. This meaning is all the more strongly associated with the affix as it can reliably be correlated with the same type of verbs and the same type of verbal variable. Obviously, morphological derivational series e.g. *chanter* ‘sing’ / *chanteur* ‘singer’, *élaguer* ‘trim’ / *élagueur* ‘trimmer’, etc. also contribute to support the soundness of the meanings in question.

It is well-known that derived nouns suffixed with *-eur* and *-oir* may also denote instruments e.g. *tondeuse* ‘clippers, shears’, *sarcloir* ‘hoe’. In this case too, the functional meaning of the derived N is correlated with the variable selected in the semantic representation of the base V. The definition of what an instrument (or an agent) is can be discussed at length (Koenig et al., 2008; Huyghe & Tribout, 2015); here I will assume that it is an object that an agent has to use to complete a given action and that this object exists before and after the action. This idea is embodied in the sketchy representation given for *tondeuse* in (12).

- (12)  $tondeuse' = \lambda z \exists x y \exists e^1 \exists e^2. [shear'(x, y, e^1) \wedge AGT(x) \wedge PAT(y) \rightarrow use'(x, z, e^2)]$

Facts such as (11)-(12) lead us to positively answer to the second question raised above: some affixes do have several meanings, among which those used to derived nouns from verbs. The

cases discussed in §1 are then of limited extension. The next step will be to assess to what extent this limitation is tied with the nature of the base (N vs. V). The presentation will show that the ontological issue unexpectedly arises anew.

Indeed the type of action that a verb denotes may affect the selection of the noun's meaning derived from this verb. For instance, the action of washing (something) does not impose to use an object dedicated to this task, which makes it implausible to derive a N denoting an instrument from this verbal meaning. On the contrary, ground-hoeing or sheep-shearing cannot be completed with bare hands, which supports the existence of derivations such as (12) (Namer & Villoing, 2008). As for meanings based on nominal ontological content, the latter generally refers to scenarios e.g. 'game~hunter' that activate verbal contents such as 'hunt'(x,y,e)  $\wedge$  AGT(x)... or to Pustejovskian qualia e.g. Origin, Users for artefacts, which provide variables to cling to. The hypothesis would be that the behavior of derivational exponents seems to strongly depend on accessibility of their base's meaning: immediate (for Vs) vs. mediate (for Ns).

As a counterpoint, the communication will also address the issue of derived nouns with a special meaning that belong to derivational series with very few attestations e.g. *chat-ière* 'catflap', of which many examples are given in Corbin & Corbin (1991).

## References

- Beavers, John. 2011. On affectedness. *Natural Language & Linguistic Theory* 29(2). 335–370.
- Beavers, John. 2013. Aspectual classes and scales of change. *Linguistics* 51. 681–706.
- Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.
- Corbin, Danielle & Pierre Corbin. 1991. Un traitement unifié du suffixe -er(e). *Lexique* (10). 61–145.
- Crysmann, Berthold & Olivier Bonami. 2015. Variable morphotactics in information-based morphology. *Journal of Linguistics* 51(1). 1–64.
- Dowty, David R. 1979. *Word meaning and montague grammar*. Dordrecht: Reidel.
- Haspelmath, Martin. 2002. *Understanding morphology*. London: Arnold.
- Haspelmath, Martin. 2020. The morph as a minimal linguistic form. *Morphology* 30(2). 117–134. doi:10.1007/s11525-020-09355-5.
- Huyghe, Richard & Delphine Tribout. 2015. Noms d'agents et noms d'instruments: le cas des déverbaux en -eur. *Langue française* 1. 99–112.
- Koenig, Jean-Pierre, Gail Mauner, Breton Bienvenue & Kathy Conklin. 2008. What with? the anatomy of a (proto)-role. *Journal of Semantics* 25(2). 175–220. doi:10.1093/jos/ffm013.
- Krifka, Manfred. 1998. The origin of telicity. In Susan Rothstein (ed.), *Events and grammar*, 197–235. Dordrecht: Kluwer Academic Publishers.
- Millikan, Ruth Garret. 2000. *On clear and confused ideas*. Cambridge M.A.: CUP.
- Namer, Fiammetta & Florence Villoing. 2008. Interpréter les noms déverbaux: quelle relation avec la structure argumentale du verbe base? In *1er congrès mondial de linguistique française*, 1539–1557. Paris.
- Rothstein, Susan. 2007. Telicity, atomicity and the vendler classification of verbs. In Susan Rothstein (ed.), *Theoretical and crosslinguistic approaches to the semantics of aspect*, 43–77. Amsterdam / Philadelphia: John Benjamins.
- Verkuyl, Henk J. 1993. *A theory of aspectuality. the interaction between temporal and atemporal structure*. Cambridge: Cambridge University Press.
- Vicente, Agustin. 2018. Polysemy and word meaning: an account of lexical meaning for different kinds of content words. *Philosophical Studies* 175(2). 00–00.

---

# The other perspective on exponence

*Sacha Beniamine*

University of Surrey,  
Surrey Morphology Group

*Mae Carroll*

Australian National University,  
School of Culture, History & Language

---

## 1 Introduction

Stump (2001, 11) defines exponence as “the only association between inflectional markings and morphosyntactic properties.”. This association is at the very core of morphology. Just as Stump’s, most definitions of exponence do not expressly indicate a directionality in the association. Yet, most morphological theories address it strictly in the direction of production, asking how marking realizes morphosyntactic meaning; or more generally how inflectional forms are produced. Production is indeed one of the core problems inflection poses to speakers. Yet, linguistic communication generally involves not only a producer, but also a receiver. Hence, a complementary task to production is that of comprehension: when hearing an inflected form, how can we recognize its morphosyntactic meaning? Which parts of words support these inferences? Surprisingly, this perspective has received much less attention. In this abstract, we describe the problems raised by a comprehension-based theory of exponence; propose a simple implemented algorithm for segmentation of fine grained formatives from inflected paradigms and a theory of exponential meaning grounded in set theory. This lets us study patterns in the discriminative power of formatives, and set the grounds for large scale typology of exponence.

## 2 The missing half of the story

Most morphological theories are formulated as accounts of production. In Hockett’s influential typology of morphological theories, all three models, item-and-arrangement, item-and-process and word-and-paradigm (Hockett, 1954), are concerned with how to produce morphological forms. Similarly, Stump’s (2001) categories along the lexical-inferential and incremental-realizational axes all apply to models of how language speakers produce inflected forms. Blevins (2016) further distinguishes constructive grammars (which focus on building words from component pieces, bottom-up) from abstractive ones (which focus on describing observed relations, top-down). Although the latter could be more suited to the comprehension perspective, work in this area has mostly focused on a productive question, called the Paradigm Cell Filling Problem (or PCFP, Ackerman et al., 2009): *How do speakers draw inferences to produce unseen forms in morphological paradigms?* An exception in this landscape is the discriminative learners of Baayen et al. (2015, 2019), which predict meaning (distributional vectors) from forms (orthographic, phonemic or acoustic) directly, without identifying any intermediate analytic units. The current work adopts the DISCRIMINATIVE perspective, and introduces the Paradigm Cell Recognition problem (or PCR, parallel to the PCFP): *Given an inflected wordform, what, in its shape, allows speakers to infer its morpho-syntactic properties?* We find that this question, which pertains to the sub-word structure, has not been addressed systematically by morphological theories.

In this perspective, we define exponence as the phenomena by which parts of words provide information about inflectional meaning. Although it may seem that the same analyses could account for both production and comprehension, we find that the pressures in favor of production and discrimination are often at odds, and can lead to different generalizations. Take for

example the small set of Latin verbal forms given in Table 1<sup>1</sup>. If we ask how to produce the passive forms from the active forms, a clear generalization emerges: add either of *-ri-*, *-er-* or *-ēr-* to the active forms. Now we ask instead a question relating to comprehension: how can one recognize that a word in this table is a passive ? A sufficient, and fully informative clue here is the presence of the *-r-*. It is not however the only informative sub-sequence: noticing the final short *-i* in *DUCO* narrows down the possible cells to either the present or the passive; and in *AMO* and *VIDEO* to either the future or the passive. Thus, *-i-* and *-r-* have different discriminative power (different exponential value), and in comprehension, should be segmented differently.

Table 1: Latin verbal subparadigms for the second person singular passive and active, present and future of a few exemplar lexemes.

Lexeme	AMO	VIDEO	DUCO	AUDIO
IND.PRS.ACT.2SG	amās	vidēs	dūcis	audīs
IND.PRS.PASS.2SG	amāris	vidēris	dūceris	audīris
IND.FUT.ACT.2SG	amābis	vidēbis	dūcēs	audiēs
IND.FUT.PASS.2SG	amāberis	vidēberis	dūcēris	audiēris

A discriminative theory of exponence departs from some of the familiar principles of economy which are relevant in production. First, in production, it is often useful to normalize un-natural distributions of exponents to more natural and compact "meanings", using mechanisms such as rules of referral Zwicky (1985) or morpheme generalization (Trommer, 2013). For the purpose of comprehension, however, the presence of a formative in a set of cells, be it natural or not, is always informative. Thus, there no need for mechanisms to fix imperfect distributions, as what may seem irregular or incoherent from the angle of production is in fact well discriminated in comprehension (Blevins et al., 2017). Second, in production, the presence of distinct formatives expressing the same morpho-syntactic properties appears as a problem, solved by postulating a set of allomorphs, associated with mechanisms (morpho-phonological constraints, inflection class indexes, etc) allowing to choose the appropriate formative for each word. In comprehension, alternants are all informative, and can be maintained as separate discriminative cues.

### 3 Discriminative exponence

In order to provide a theory of exponence, we are faced with two methodological problems: that of segmentation and of meaning assignment (Manning, 1998; Spencer, 2012; Trommer, 2013):

- **Segmentation problem:** Given a set of inflected word-forms belonging to the same paradigm, what are the minimally informative sub-sequences (or formatives) ?
- **Meaning Assignment problem:** Given the structured morphological paradigm of a lexeme, and a segmentation of its word forms into formatives, what grammatical information can each formative provide ?

We provide a solution to both, starting from fully inflected (but unsegmented) paradigm forms, organized in paradigm tables, with each cell associated with a set of morpho-syntactic

<sup>1</sup>We thank Olivier Bonami for suggesting this particular example.

properties <sup>2</sup>.

We start with segmentation, isolating sequences which systematically co-vary in the paradigm. We call these *FORMATIVES* (after Pike, 1963) defined as the longest contiguous substrings which always recur together in a paradigm. Formatives are identified on the basis of the set of cells in which they recur, observed in matrices of automatically aligned forms (our method for alignment refines that of Beniamine & Guzmán Naranjo, 2021). The operation is first done separately on each phonological tier, in order to allow for supra-segmental exponence. We call *distribution* the set of cells in which a formative occurs. Any formative present in all cells are identified as constant, inert stems (following Bonami & Beniamine, 2021). We focus on formatives which occur in proper subsets of the paradigm cells. Table 2 shows the alignments for the Latin forms of *AMO* in Table 1, and the distributions of each formative segment. Note that this segmentation does not coincide with traditional segmentations in stems and exponents. For example, in the case of suppletion, only the common substring to all stems (if any) will be considered inert. Any substring which occurs in a proper sub-set of the cells is informative in discrimination, and thus has exponential power (This is not to say that may not also contribute lexical information).

Table 2: Alignments and distributions of formatives for a small Latin sub-paradigm of *AMO*

IND.PRS.ACT.2SG	ama:	-	-	-	-	s				
IND.PRS.PASS.2SG	ama:	-	-	r	i	s				
IND.FUT.ACT.2SG	ama:	b	-	-	i	s				
IND.FUT.PASS.2SG	ama:	b	e	r	i	s				
	slots	0	1	2	3	4	5			

			ACT	PASS				ACT	PASS
PRS									
FUT	b	b						e	

			ACT	PASS				ACT	PASS
PRS			r					i	
FUT			r				i	i	

This segmentation procedure produces pairs of formatives and their distribution, eg.  $\langle /r/, \{ \{ \text{IND}, \text{PRS}, \text{PASS}, \text{2SG} \}, \{ \text{IND}, \text{FUT}, \text{PASS}, \text{2SG} \} \} \rangle$ . These distributions, stated as sets of feature value sets, form part of the knowledge that a language user can be said to have about a given formative. We provide set-theoretic definitions for formatives and distributions, and show how to derive generalized and accurate definitions of their meaning. This relies on defining the set of all minimal and informative descriptions of the distribution. The association between a formative and this morphosyntactic meaning forms the exponence relationship. In the small sub-paradigm of *AMO* from Table 1, the information carried by  $/r/$ ,  $\text{exp}(/r/)$  is  $\{ \text{PASS} \}$  (as it is present exactly in the set of cells described by the feature passive, nothing less, nothing more), whereas  $\text{exp}(/i/) = \{ \text{FUT}, \text{PASS} \}$  (as it is present both in all the future cells, and all the passive cells, nothing less, nothing more).

## 4 Conclusion

We describe a formal theory of discriminative exponence, which describes how information for the comprehension task is organized and distributed in inflected words. Both the procedure for segmentation and meaning assignment are entirely formalized and implemented, and can be applied to machine readable lexicons of inflected forms. This methodology lets us produce com-

<sup>2</sup>we recognize that such tables are themselves the result of analyses. Within this work, they are taken as axioms, although our theory can be applied to different paradigmatic analyses of the same data, with potential for meta-theoretical comparisons.

parable analyses of typologically different inflectional systems. On this basis, observe patterns of exponence in verbal lexicons of Modern Standard Arabic verbs, French, Georgian, Navajo, Ngkolmpu, and Yaitepec Chatino. We show how our analyses capture generalizations about the discriminative nature of exponents, and produce clear, comparable and reproducible accounts of phenomena of interest in the typology of exponence, such as inflection classes, cumulation, syncretism, verbosity, etc. For the purpose of this presentation, we draw analyses from entire lexicons. We leave to future work the study of analytic variation when sampling sub-sets of forms, as well as the treatment of reduplication and metathesis.

## References

- Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar*, 54–82. Oxford: Oxford University Press.
- Baayen, R. H., C. Shaoul, J. Willits & M. Ramscar. 2015. Comprehension without segmentation: A proof of concept with naive discrimination learning. doi:10.1080/23273798.2015.1065336.
- Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 1–39. doi:10.1155/2019/4895891.
- Beniamine, Sacha & Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. *Proceedings of the Society for Computation in Linguistics* 4(21). <https://scholarworks.umass.edu/scil/vol4/iss1/21>.
- Blevins, James P. 2016. *Word and Paradigm Morphology*. Oxford: Oxford University Press.
- Blevins, James P., Petar Milin & Michael Ramscar. 2017. The Zipfian Paradigm Cell Filling Problem. In Ferenc Kiefer, James P. Blevins & Huba Bartos (eds.), *Morphological paradigms and functions*, 141–158. Leiden: Brill.
- Bonami, Olivier & Sacha Beniamine. 2021. Leaving the stem by itself. In Marcia Haag, Sedigheh Moradi, Andrija Petrovic & Janie Rees-Miller (eds.), *All things morphology: Its independence and its interfaces*, vol. 353 Current Issues in Linguistic Theory, 81–98. Amsterdam: John Benjamins Publishing Company. doi:10.1075/cilt.353.05bon.
- Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10. 210–234.
- Manning, Christopher D. 1998. The segmentation problem in morphology learning. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning* NeMLaP3/CoNLL '98, 299–305. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1603899.1603953.
- Pike, Kenneth L. 1963. Theoretical implications of matrix permutation in fore (new guinea). *Anthropological Linguistics* 5(8). 1–23. <http://www.jstor.org/stable/30022433>.
- Spencer, Andrew. 2012. Identifying stems. *Word Structure* 5(1). 88–108. doi:10.3366/word.2012.0021.
- Stump, Gregory T. 2001. *Inflectional morphology. a theory of paradigm structure*. Cambridge: Cambridge University Press.
- Trommer, Jochen. 2013. Paradigmatic generalization of morphemes. *Morphology* 23(2). 269–289. doi:10.1007/s11525-013-9226-4.
- Zwicky, Arnold M. 1985. How to describe inflection. *Annual Meeting of the Berkeley Linguistics Society* 11. doi:10.3765/bls.v11i0.1897.



---

# The boundaries of person: parameters of variation

*Grigorij Sibiljof*

Russian State University for the  
Humanities

---

## 1 Introduction

The position of third person pronoun among other personal pronouns was discussed in numerous papers, including those by Jespersen (1924), Lyons (1977), Cysouw (1997), and Bhat (2004). The later (Bhat 2004: 134-138) put forward an idea of two types of pronominal systems differentiated based on whether or not the third person pronoun should be included in the personal pronoun system. In the first type the third person pronoun is related to demonstratives, whereas in the second type it is not. Languages thereby can be classified into ‘two-person languages’ and ‘three-person languages’. This idea, however, was never fully appreciated. We believe that it could be turned into a typological classification of pronominal paradigm mapping in the languages of the world that could, apart from just classifying languages into types, reflect or even predict some diachronic changes within their pronominal systems.

## 2 Parameters of variation

For the classification to work properly, we suggest three parameters of variation that are independent of one another and typologically relevant.

The first parameter (‘the lexical parameter’) is a revised version of Bhat’s (2013) relatedness to demonstrative pronouns. This parameter differentiates whether the third person pronoun shares a stem with a neighbouring non-personal category. Unlike Bhat (2013) we propose that not only demonstratives be considered in this parameter, but only stem sharing should be relevant.

The other two parameters take into account morphological structure of the pronominal paradigm. The first one, the ‘strong paradigmatic parameter’, characterizes the system as to the number and nomenclature of the morphological features carried by the third person pronoun compared both to other personal pronouns and any related category. This parameter includes one of Bhat’s (2013) values, namely the relatedness by gender markers, but we suppose that lexical relatedness should be considered as orthogonal to the paradigm organization.

The ‘weak paradigmatic parameter’ differentiates between systems where the 3rd person pronoun exhibits differences in morphological pattern in the shared features from those systems where it does not.

## 3 Typological relevance

All three parameters are typologically relevant and independent of one another.

For example, Nivkh and Buryat pronominal system only differ by the lexical parameter. In Nivkh (Gruzdeva 1998) 3rd person pronouns *if* ‘3sg’ and *imŋ* ‘3pl’ are not lexically related to demonstratives set *-yd* (or any other categories). Opposite to Nivkh’s system, in Buryat (Skribnik 2003) demonstratives *ene* ‘this’ and *tere* ‘that’ are used as 3rd person pronouns. Both languages have the same value of both paradigmatic parameters, namely, the paradigm

of the 3rd person pronoun in both languages consists of exactly the same features and values and shows exactly the same morphological patterns as the 1st and 2nd person pronouns.

The systems in Finnish and Puyuma only differ by the value of the strong paradigmatic parameter. In Finnish pronominal system (Karlsson 1999) 3rd person pronoun *hän* is not related to demonstrative system in any way (or any other categories) and distinguishes all features that 1st & 2nd person pronouns do. It contrasts with Puyuma system (Deng 2018) where 3rd person pronoun *tu* is not related to any other categories either, but differs from other personal pronouns in that it lacks plurality distinction.

The differences between the pronominal systems of Lango and Yukagir base only on the weak paradigmatic parameter. In Lango (Noonan 1992) 3rd person pronoun is not related to any other category and shares all the morphological features with the 1st & 2nd personal pronouns. On the other hand, in Yukagir (Maslova 1993) 3rd person pronoun, not related to any other category and sharing all the features with the 1st and 2nd person pronouns, has a truncated case system (no predicative or pronominal accusative compared to the rest of personal pronouns), therefore differing from the other personal pronouns by the organization of the paradigm.

#### 4 Values for the suggested parameters

First, the lexical parameter has two base values: one is ‘no lexical relatedness to any neighbouring category’ and the other ‘lexical relatedness to some category’, with further division to subvalues determining the category 3rd person pronoun is related to (demonstrative, nominal classifier etc.). The paradigmatic parameters have quantity value set (the number of differences the systems exhibit) and quality set (what kind of differences those are) with the later set being non-exclusive.

The parameters discussed above have been tested on a sample of 40 languages and listed below are some values for the proposed parameters that we have found so far.

For the weak parameter we can only suggest the two obvious quantitative values: (i) 3rd person pronoun paradigm shows different patterns than some related category and/or 1st & 2nd pronouns, (ii) paradigmatic structure of 3rd person pronoun is the same as the related category and/or 1st & 2nd pronouns.

For the strong paradigmatic parameter four major quality values have been found, namely those in (i) gender, (ii) animacy, (iii) plurality and (iv) case system. In our sample the quantity value set for this parameter varies from 0 (3rd person pronoun shares all the morphological features with any related category and/or 1st & 2nd pronouns) to 2 (e.g. in Guarani (Gregores & Suárez 1933) 1st & 2nd pronouns inflect for case and number whereas 3rd person pronoun does not).

Interestingly, it seems quite typical for 3rd pronoun and the demonstratives to distinguish gender especially in comparison to the 1st and 2nd person pronouns. Seven languages out of our sample show difference in gender distinction between 1st & 2nd person pronouns and 3rd person pronouns, while other related categories align with 3rd person pronoun in this feature (the system of Jacalteco (Craig 1977), where only 3rd person pronoun have gender category, is an exception for that).

Languages with the animacy differences between the categories (like Finnish (Karlsson 1997) where 3sg pronoun *hän* can only be used for animate objects and demonstrative *se* is used elsewhere) show that the animate objects may be considered as more typically ‘third-person’ than inanimate.

As to the quantitative value set, we can suggest that systems with more than two differences in features are highly unlikely (it has not been found so far), while languages with one or two differences in the set of features occupy roughly one third of our sample.

The lexical parameter differentiates five types of systems: (i) 3rd person pronoun is unrelated to any other category, (ii) the demonstratives are used instead of 3rd person pronoun (see Buryat pronominal system above), (iii) 3rd person pronoun share stem with distal demonstrative (for example, Turkish (Lewis 1967) distal demonstrative *o* functions as 3rd person pronoun), (iv) 3rd person pronoun share stem with medial demonstrative (in Mapuche (Smeets 1989) *fey* is a medial demonstrative used to refer to 3rd person), (v) 3rd person pronoun share stem with special anaphoric demonstrative (cf. Lezgian anaphoric demonstrative *ha* used in pronominal contexts (Haspelmath 1993)), (vi) 3rd person pronoun is lexically related to some non-demonstrative category (nominal classifier in Jacaltec (Craig 1977)).

## 5 Discussion

We do not argue that the values above form an exhaustive list of possible values. For example, in Yami (Rau 2006) which is not presented in our sample the proximal demonstrative *iya* functions as 3rd person pronoun, which does not fall into any of the types considered above.

However, we do argue that the classification presented above is predictive of the language changes that occur in this part of the language system. The lexical parameter presumably shows the boundary where the transition between the 3rd person pronoun and some neighbouring category has already happened. The paradigmatic parameters show how much morphological difference exists on the boundary between two neighbouring categories and whether or not it is crossable. Therefore, with the paradigmatic parameter at hand we can predict the possible diachronic transition between the categories.

Moreover, with the suggested parameters combined there are more than just two types of languages, the ‘two-person’ and ‘third-person’ systems offered by Bhat (Bhat 2004). We claim that there is at least one other type, the ‘three-part system’, where the third person pronoun neither has lexical relatedness to other categories, nor can it be included in the personal pronoun system due to considerable differences in morphological features.

## References

- Bhat, D.N.S. 2004. Pronouns. Oxford University Press.
- Bhat, D.N.S. 2013. Third Person Pronouns and Demonstratives. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/43>, Accessed on 2022-09-03.)
- Craig, Colette Grinevald. 1977. The Structure of Jacaltec. Austin: University of Texas Press.
- Cysouw, Michael. 1997. 3rd Person Personal Pronoun: a Universal Category? Unpublished Manuscript (Available via link <https://cysouw.de/home/manuscripts.html>).
- Deng, Fangqing. 2018. Binanyu yufa gailun (General grammar of Puyuma) – Second edition – Austronesian Languages of Taiwan Series, 13. - Taipei: Yuanliu.
- Gregores, Emma and Suárez, Jorge A. 1967. A Description of Colloquial Guaraní. The Hague: Mouton.
- Gruzdeva, Ekaterina. 1998. Nivkh. (Languages of the World/Materials, 111.) München and Newcastle: Lincom Europa.

- Haspelmath, Martin. 1993. *A Grammar of Lezgian*. (Mouton Grammar Library, 9.) Berlin: Mouton de Gruyter.
- Jespersen, O. (1924) *The Philosophy of Grammar*, George Allen & Unwin, London.
- Karlssohn, Fred. 1999. *Finnish: an essential grammar*. London: Routledge. (Translated by Andrew Chesterman).
- Lewis, Geoffrey L. 1967. *Turkish Grammar*. Oxford: Clarendon Press.
- Lyons, J. (1977) *Semantics*, Cambridge University Press, Cambridge.
- Maslova, Elena. 1999. *A Grammar of Kolyma Yukaghir*. (Published as Maslova 2003a).
- Noonan, Michael. 1992. *A Grammar of Lango*. (Mouton Grammar Library, 7.) Berlin: Mouton de Gruyter.
- Rau, Victoria and Maa-Neu Dong. 2006. *Yami texts with a reference grammar and dictionary*. – Taipei: Language and linguistics monograph series, A10
- Skribnik, Elena. 2003. Buryat. In Janhunen, Juha (ed.), *The Mongolic Languages*, 102-128. London: Routledge.
- Smeets, Ineke. 1989. *A Mapuche Grammar*.

---

# Back-formation, forward-formation, and cross-formation in the same construction

## The case of Hungarian compound verbs

László Palágyi

Eötvös Loránd University of Budapest

---

### 1 Introduction

Hungarian compound verbs have become more and more productive in the last few decades, analyses have shed light on the great number of hapaxes as well as conventional verbs that represent this pattern (Pusztai 1999; Ladányi 2017: 647). Hungarian compound verbs have traditionally been considered as the outcomes of  $V \leftarrow N$  back-formation (Lengyel 2000: 344–345), created by removing an affix (typically the nominalizer suffix *-ás/-és*), as illustrated below.

- |     |                         |   |                                       |
|-----|-------------------------|---|---------------------------------------|
| (1) | utas-tájékoztat-ás      | → | utas-tájékoztat <sup>1</sup>          |
|     | passenger-inform-NLZR   |   | passenger-inform                      |
|     | ‘passenger information’ |   | ‘provide passengers with information’ |
|     |                         |   | (lit. ~ ‘to passenger-inform’)        |

Productivity of this pattern challenges the view that back-formation has only diachronic relevance (see Marchand 1969). However, Hungarian back-formation cannot be considered a reversal of a word-formation rule either (cf. Aronoff 1976) because not only compound nouns containing a productive nominalizer suffix can function as input. Some studies see morphological reanalysis behind back-formation in Hungarian (Ladányi 2017: 646; cf. Mel’čuk 2001: 532). Under this approach, the above mentioned compound verb *utastájékoztat* ‘provide passengers with information’, back-formed from the compound noun *utastájékoztatás* ‘passenger information’, requires a process in which the head-modifier structure becomes a head-complement structure. According to this explanation, the reanalysis illustrated below makes the removal of the nominalizer *-ás* possible.

- |     |   |   |   |
|-----|---|---|---|
| (2) | [[utas] <sub>N</sub> tájékoztat-ás] <sub>N</sub>    | → | [[utas-tájékoztat] <sub>V</sub> ás] <sub>N</sub>    |
|     | [[passenger] <sub>N</sub> inform-NLZR] <sub>N</sub> |   | [[passenger-inform] <sub>V</sub> NLZR] <sub>N</sub> |

The present research on Hungarian compound verbs argues that appeals to back-formation as a particular morphological process (cf. Bauer 1983: 232; Lieber 2005: 375; Štekauer 2015) only scratches the surface of a phenomenon whose formal realization is of secondary importance. It rejects the notion of morphological reanalysis and demonstrates that the same associative and analogical relations might result in back-formation, forward-formation, and cross-formation. This diversity raises the issue of generalizations in derivation, and motivates a typology of derivational processes according to which Hungarian compound verbs can be characterized.

---

<sup>1</sup> All examples are from HNC (Hungarian National Corpus) (Oravecz et al. 2014)

## 2 Findings

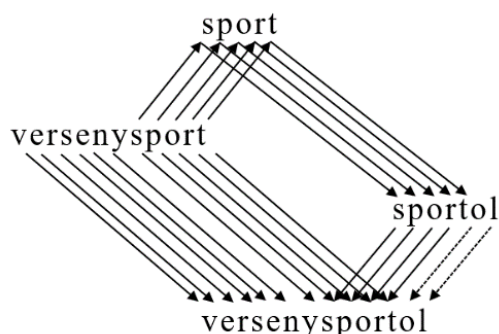
Previous research on the field did not detect that Hungarian compound verbs can arise not only by back-formation. They can be created in rather diverse ways, as shown by (3a) and (3b). Example (3a) exhibits forward-formation, (3b) exhibits cross-formation.

- |        |  |   |   |
|--------|--|---|---|
| (3) a. | verseny-sport<br>competition-sport<br>'competitive sport'        | → | verseny-sport-ol<br>competition-sport-VLZR<br>'do competitive sport'<br>(lit. to competitive-sport) |
| b.     | gén-manipul-áció<br>gene-manipul-ation<br>'genetic modification' | → | gén-manipul-ál<br>gene-manipul-ate<br>'modify genetically'<br>(lit. to gene-manipulate)             |

The Hungarian compound verbs *versenysportol* and *génmanipulál* are problematic for the account based on morphological reanalysis as the head component of the nominal compound should be a zero morpheme (4a) and a segment subjected to deletion (4b), respectively.

- |        |   |   |   |
|--------|---|---|---|
| (4) a. | *[[verseny] <sub>N</sub> sport-Ø] <sub>N</sub><br>[[competition] <sub>N</sub> sport-Ø] <sub>N</sub> | → | [verseny-sport] <sub>V</sub> Ø] <sub>N</sub><br>[[competition sport] <sub>V</sub> Ø] <sub>N</sub> |
| b.     | *[[gén] <sub>N</sub> manipul-áció] <sub>N</sub><br>[[gene] <sub>N</sub> manipul-ation] <sub>N</sub> | → | [[génmanipul] <sub>V</sub> áció] <sub>N</sub><br>[[gene manipul] <sub>V</sub> ate] <sub>N</sub>   |

As for forward-formation in example (3a), it is remarkable that the productivity of the verbalizer suffix *-(V)l* does not per se explain suffixation, as this suffix is typically added to foreign words and monosyllabic nouns (Ladányi 2017: 555). However, the existing simple verb *sport-ol* 'do sport' derived from *sport* 'sport' might explain the use of suffix *-(V)l* in *verseny-sport-ol* 'do competitive sport'. The word-based (see Blevins 2006) associative network behind the compound verb in question is illustrated below. The nominal head of *verseny-sport* 'competitive sport', i.e. *sport* 'sport' evokes the derivative *sport-ol* 'do sport', which is why *verseny-sport-ol* 'do competitive sport' emerges via forward-formation.



**Figure 1 Associative network underlying *verseny-sport-ol* (competition-sport-VLZR) 'do competitive sport'**

In fact, the verb associated with the head of the compound noun predicts the form of the verbal head in all cases. The diversity of morphological processes can be traced back to the diversity of noun $\leftrightarrow$ verb relationships in Hungarian, where the evoked verb can be simple as well as denominal. The nominal head of *utastájékoztatás* ‘passenger information’ (see example (1)) is *tájékoztatás* ‘(providing people with) information’, which evokes the base verb *tájékoztat* ‘inform’. This is why *utastájékoztat* ‘provide passengers with information’ emerges via back-formation. By the same token, the nominal head of *génmanipuláció* ‘genetic manipulation’ is *manipuláció* ‘manipulation’, which evokes its sister verb *manipulál* ‘manipulate’, and motivates the emergence of *génmanipulál* ‘modify genetically’ via cross-formation.

It is remarkable that the morphological transparency of compound verbs is based on the associative network of words and not that of morphemes. A morpheme-based approach can hardly posit a plausible rule that might be accompanied by affix deletion (cf. (1)), affixation (cf. (3a)), and affix substitution (3b) alike. Even if the structure of the derivatives is highly varied, Hungarian compound verbs represent the same productive way of derivation that can be described as an analogical process (with analogy considered here as a domain-general cognitive process responsible for productivity, see Bybee 2010). As it is shown below, four lexical clusters contribute to the formation of compound verbs in Hungarian. A simple noun relates to a simple verb just like a compound noun to a compound verb. The principles of mapping are as follows: a simple noun is to a simple verb as a compound noun is to a compound verb. This analogical operation may trigger back-formation, forward-formation, and cross-formation.

Simple nouns	:	Simple verbs	~	Compound nouns	:	Compound verbs
N	:	N-VLZR	~	N-N	:	(FF) <sup>2</sup> N-N-VLZR
sport		sport-ol		verseny-sport		verseny-sport-ol
V-NLZR	:	V	~	N-V-NLZR	:	(BF) N-V
tájékoztat-ás		tájékoztat		utas-tájékoztat-ás		utas-tájékoztat
X-NLZR	:	X-VLZR	~	N-X-NLZR	:	(CF) N-X-VLZR
manipul-áció		manipul-ál		gén-manipul-áció		gén-manipul-ál

Figure 2 Analogical relations motivating the formation of compound verbs

### 3 Discussion

The investigation of Hungarian compound verbs leads us to the issue of generalizations in word-formation. It is crucial to describe how Hungarian compound verbs can be treated as instances of the same morphological pattern.

In order to handle the word-based nature of the pattern, we consider the Hungarian compound verb as a construction, i.e., a systematic pairing of meaning and form (cf., Booij and Audring 2017). We assume that derivational constructions are based on source-oriented and product-oriented generalizations (cf., Bybee 2001; Kapatsinsky 2013). Source-oriented generalizations are based on the associative relationship between distinct constructions, they involve information about the scope (word-class, phonological structure, lexical group, etc.)

---

<sup>2</sup> In Figure 2, “FF”, “BF”, and “CF” are used as abbreviations for “forward-formation”, “back-formation”, and “cross-formation”.

and the way of (analogical) mapping between the base and the derivative. Product-oriented generalizations, for their part, provide information about the schematic meaning and form of a construction. These types of generalizations are obviously not mutually exclusive, they are typically interrelated in most derivational constructions. However, derivational constructions also vary in the relative prominence of source-oriented and product-oriented generalizations. For instance, diminutive constructions and many onomatopoeic verbal constructions in Hungarian can only be characterized by product-oriented formal generalizations. The former do not provide any information about the way of mapping, and the latter do not define any scope (which means that the derivatives do not have bases). In terms of this classification of constructions, Hungarian compound verbs represent the other end of the scale. They can be characterized by purely source-oriented formal generalizations that concern the scope and the way of mapping (see Figure 1 and 2). This conclusion partly confronts the notion that “many, if not all, schemas are product-oriented rather than source-oriented” (Bybee 2001: 128).

The account outlined above might have theoretical implications for the analysis of back-formation in other languages as well. It can be hypothesized, for example, that back-formation depends on schemas for which mainly source-oriented formal generalizations are responsible.

## 4 References

- Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Cambridge: The MIT Press.
- Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.
- Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42. 531–573.
- Booij, Geert & Jenny Audring. 2017. Construction Morphology and the Parallel Architecture of Grammar. *Cognitive Science* 41(S2). 277–302.
- Bybee, Joan 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
- Bybee, Joan 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Kapatsinski, Vsevolod. 2013. Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language* 89(1). 110–148.
- Ladányi, Mária. 2017. Alaktan. [Morphology] In Tolcsvai Nagy, Gábor (ed.), *Nyelvtan* [Grammar], Budapest: Osiris Kiadó. 503–660.
- Lengyel, Klára. 2000. Ritkább szóalkotási módok. [Less common processes of word-formation] In Keszler Borbála (ed.), 2000. *Magyar grammatika* [Hungarian Grammar], Budapest: Nemzeti Tankönyvkiadó.
- Lieber, Rochelle. 2005. English word-formation processes. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word formation*. Dordrecht: Springer. 375–427
- Marchand, Hans 1969. *The Categories and Types of Present-day English Word-formation a Synchronic-Diachronic Approach*. München: Verlag C.H. Beck.
- Mel’čuk, Igor. 2001. Formal processes. In Geert Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphologie / Morphology (vol 1)*. Berlin / New York: de Gruyter. 523–535
- HNC ~ Oravec, Csaba & Váradi, Tamás & Sass, Bálint. 2014. The Hungarian Gigaword Corpus. *Proceedings of LREC 2014*. 1719–1723.
- Pusztai, Ferenc. 1999. A szóalakutánzás elve és kategóriája. [Principle and category of wordform-imitation]. Kugler, Nóra & Lengyel, Klára (eds.), *Ember és nyelv*. [Human and language] Budapest: ELTE BTK Mai Magyar Nyelvi Tanszék, 266–271.
- Štekauer, Pavol. 2015. Back-formation. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation. An international handbook of the languages of Europe (vol 1)*. de Gruyter, Berlin / New York. 340–352.



---

# Nouns in *-ion* and denominal verbs: can the output be explained?

## The case of English and French

Aurélie Héois

Centre d'Études Linguistiques – Corpus,  
Discours et Sociétés

---

### 1 Introduction

Many studies on denominal verbs, both in French and English, have focused their analyses on the main word-formation process involved, i.e. conversion (which is by far the most productive pattern), as exemplified in Bleotu 2019; Clark & Clark, 1979; Kiparsky 1997, among others. Some tackle cases of conversion and suffixation (Huyghe, 2017; Plag 1999) while studies on prefixation are relatively scarce (Jacquey & Namer 2007)<sup>1</sup>. Similarly, backformation is often forgotten in these accounts on denominal verbs even though Nagano (2007: 67-68) shows that noun-to-verb and adjective-to-verb backformations are the only two productive patterns of backformation in contemporary English. Nagano (2007) offers an original account of backformation as a type of conversion, revisiting Marchand's hypothesis (1960; 1969) according to which backformation is a combination of zero derivation – to which Nagano prefers conversion – and clipping – which is unpredictable.

In this study, I would like to confront data on denominal verbs to one crucial assumption made in Nagano (2007: 59):

*When an input has a nominal or adjectival (pseudo-) suffix (e.g., television<sub>N</sub>), conversion to a verb yields a categorially verbal but formally nominal/adjectival output (e.g. television<sub>V</sub>). In such cases, conversion uses clipping to remove the categorially obstructive element, i.e., the (supposed) suffix, to adjust the output form to the output category.<sup>2</sup>*

In other words, language users try to avoid ambiguity: if a converted verbal ending clearly belongs to another category, Nagano's assumption predicts that conversion will be completed with clipping so as to delete the “obstructive element”.

To question this assumption, I selected denominal verbs built on nouns ending in *-ion* in French and English. The *Oxford English Dictionary online*, along with the *Trésor de la Langue Française informatisé*, lists *-ion* as an allomorph of the *-tion* suffix. The latter is a nominalization suffix inherited from Latin in both languages with the ability to form action nouns, usually on a verbal base. In the data I collected, most cases<sup>3</sup> of backformations are of the type BF\_?ion where “?ion” stands for any deleted material ending in *-ion*.

### 2 Data

I adopt here a data-driven view and the perspective is “polysynchronic”: I consider each attested denominal verb at its time of coinage so as to analyse the patterns in denominal verbal word-formation processes. Consequently, I adopt an onomasiological approach as I place my

---

<sup>1</sup> Studies on *-ize* or *-iser* suffixation are also frequent but generally focus on the morphological paradigm and as such encompass denominal as well as deadjectival verbs (i.e., Lignon 2013).

<sup>2</sup> My emphasis.

<sup>3</sup> Respectively 50% of French backformations and 42% of the English data are of the *\_?ion* type.

analysis at the hypothetical time an innovator creates a verb from a noun in order to convey a certain meaning.

The present study is based on the data collected for an ongoing project on denominal verbs called “Vdenom”. This section briefly presents the methodology used to collect both the VdenomEN and VdenomFR data from which the present data is extracted.

## 2.1 Collecting the data

The extraction of denominal verbs was carried out manually in four online dictionaries and one paper dictionary: namely the *Oxford English Dictionary online* and *Green’s Dictionary of Slang*, for the English data, and the *Grand Robert*, the *Trésor de la Langue Française informatisé* and the *Dictionnaire historique et philologique du français non conventionnel* (Enckell 2017) for French. The resources were evaluated according to the presence of the following criteria: date of first attestation (after 1800); etymology (denominal verbs); historical data (first attested meaning). Data collection resulted in a set of 5 932 English denominal verbs attested from 1800 onwards, and 2 368 French denominal verbs coined during the same period. In the case of polysemous verbs, only the first attested meaning was selected.

## 2.2 Extracting the data

To explore Nagano’s assumption, I randomly selected a sample of 600<sup>4</sup> verbs in each dataset. In the two samples, conversion is by far the most represented word-formation process accounting for 63% of the English sample and 57% of the French sample. Backformation is the second most frequent process for English (15%) while suffixation is a close third (12%). French shows a different structure as suffixation (13,5%), prefixation (13%) and backformation (12%) are all quite equally represented.

## 3 Data description

By applying a textual filter to the base-noun column, I selected, in both samples, all verbs deriving from a noun ending in *-ion*. This results in two subsets presented in Tables 1 and 2.

### 3.1 English verbs

Table 1 shows that 52 verbs, out of the 600 English sample, correspond to the criterion “base-noun ending in *-ion*”. Three-quarter of them are cases of backformation while the last quarter is either a conversion (for half) or another process. The data tends to confirm Nagano’s assumption that an apparent nominal suffix, such as *-ion*, will rarely be kept through verbal derivation. A closer look at the cases of conversion is however needed to understand why these cases “resisted” backformation.

Morpho1_V <sup>5</sup>	Morpho2_V	Count	Examples N > V
BF	BF_ion	26 (50%)	convection > convect
BF	BF_ation	10 (19.23%)	excystation > excyst
BF	BF_tion	2 (3.85%)	sorption > sorb

<sup>4</sup> The choice of a 600 threshold, while partly arbitrary, is driven by the observation of the data and its structure through the modelling process.

<sup>5</sup> In the columns “Morpho1\_V” and “Morpho2\_V”, which both describe the morphology of the verb, “BF” stands for “backformation”, “CONV” for “conversion”, “CPX” for “complex processes”, “PREF” for “prefixation” and “SUFF” for “suffixation”.

BF	BF_ation	2 (3.85%)	relexification > relexify
CONV	CONV	6 (11.54%)	pincushion > pincushion
CPX	BF_ion/SUFF_fy	1 (1.92%)	destruction > destructify
CPX	BF_ition/SUFF_ize	1 (1.92%)	premunition > premunize
CPX	PREF_de/SUFF_ize	1 (1.92%)	ion > deionize
PREF	PREF_pre	1 (1.92%)	tension > pre-tension
SUFF	SUFF_ize	2 (3.85%)	salvation > salvationize
<b>TOTAL</b>		<b>52 (100%)</b>	

Table 1. Morphology of English denominal verbs built on N-*ion*

### 3.2 French verbs

The situation is quite different for the French data as conversion accounts for almost a third of the 56 verbs of the sample, while backformation remains the main process (66%). This suggests that Nagano’s assumption may not apply in the same way to languages other than English: either some language specific variables are needed to explain this difference, or backformation is motivated by other variables.

Morpho1_V	Morpho2_V	Count	Examples
BF	BF_ation	28 (50%)	hominisation > hominiser
BF	BF_ion	7 (12.50%)	péréquation > péréquater
BF	BF_tion	1 (1.79%)	involution > involuer
BF	BF_ction	1 (1.79%)	transduction > transduire
CONV	CONV	16 (28.57%)	fusion > fusionner
SUFF	SUFF_iser	2 (3.57%)	ion > ioniser
SUFF	SUFF_aliser	1 (1.79%)	administration > administrationaliser
<b>TOTAL</b>		<b>56 (100%)</b>	

Table 2. Morphology of French denominal verbs built on N-*ion*

## 4 Discussion

The English data suggests that Nagano’s assumption may be right, and that the main motivation for backformation may be to erase what could be construed as a categorial suffix other than verbal, and thus avoid ambiguity. Indeed, *-ion* and its allomorphs are clear cases of nominal suffixes. Still, the few cases of conversion in the sample suggest that other variables may play a role as backformation potential blockers, as illustrated in examples (1) to (3):

(1) lion<sub>v</sub>: “to frighten, to intimidate” [GDS 2022] < lion<sub>N</sub> → prosodic blocking

(2) accordion<sub>v</sub>: “To cause (a thing) to fold, collapse, etc.” [OED 2023] < accordion<sub>N</sub> → conceptual blocking

(3) proposition<sub>v</sub>: “To propose sexual activity, esp. of a casual or illicit nature, to (a person)” [OED] < proposition<sub>N</sub> → paradigmatic blocking

Another problem is raised with the French data: even though backformation remains the main process for this type of verbs, conversion accounts for almost a third of the data. Denasalization could easily explain this difference as it regularly applies to verbal conversion when the base-noun ends in a nasal vowel. This phonological process renders Nagano’s argument void for the French data as there is no longer any ambiguity between the noun and the verb. As a result, the overwhelming presence of backformation in the French sample becomes surprising. This would suggest that backformation is indeed attracted to suffix-like endings whether or not ambiguity may occur. The motivation for backformation needs to be found elsewhere, however.

Based on these two datasets, I propose that noun prosody can either block backformation (for mono- and disyllabic base-nouns) or encourage it (for base-nouns of 5 to 7 syllables). Moreover, the data suggests that paradigmatic criteria tend to directly influence the derivation. The preexistence of a paradigm in the lexicon, such as {-ation<sub>N</sub>/ -ate<sub>V</sub>} or {-isation<sub>N</sub>/ -iser<sub>V</sub>}, will pave the way for a specific type of derivation. Paradigmatic influence can also apply on a smaller scale as in the pairs {destruction<sub>N</sub>/ destruct<sub>V</sub>}// {construction<sub>N</sub>/ construct<sub>V</sub>} or {re-revolution<sub>N</sub>/ re-revolutionize<sub>V</sub>}// {revolution<sub>N</sub>/ revolutionize<sub>V</sub>}. It appears, however, that paradigmatic influence can also have the reverse effect as is the case for (3) {proposition<sub>N</sub>/ proposition<sub>V</sub>}# {-position<sub>N</sub>/ -pose<sub>V</sub>}. This example suggests, in agreement with Lignon & Namer (2014: 13), that conversion is preferred here because of the semantic specialization of the noun and subsequent verb.

## References

- Bleotu, Adina Camelia. 2019. *Towards a Theory of Denominals: A Look at Incorporation, Phrasal Spell-Out and Spanning*. Leiden: Brill.
- Clark, Eve. V. & Clark Herbert H. 1979. When Nouns Surface as Verbs. *Language*, 55(4), 767-811.
- Enckell, Pierre. 2017. *Dictionnaire historique et philologique du français non conventionnel*. Paris : Classiques Garnier Numérique.
- Green, Jonathon. 2022. *Green's Dictionary of Slang - Digital Edition*. [greensdictofslang.com](https://greensdictofslang.com)
- Dictionnaires le Robert. 2022. *Le Grand Robert - Version numérique*. [grandrobert.lerobert.com](https://grandrobert.lerobert.com)
- Huyghe, Richard. 2017. Les verbes dérivés de noms de matière. *Le Français Moderne*, 85, 208-232.
- Jacquey, Evelyne & Namer Fiammetta. 2007. Morphosémantique et modélisation : Les verbes dénominaux préfixés par é-. In Denis Bouchard, Ivan Evrard & Etleva Vocaž (éds.), *Représentation du sens linguistique : Actes du colloque international de Montréal*. 53-68. Bruxelles : De Boeck.
- Kiparsky, Paul. 1997. Remarks on Denominal Verbs. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex Predicates*. 473-499. Stanford: CSLI Publications.
- Lignon, Stéphanie. 2013. -iser and -ifier suffixations in French: Verifying data to 'verize' hypotheses? In Nabil Hathout, Fabio Montermini & Jesse Tseng (eds.), *Selected proceedings of Décembrettes 7*. München: Lincom Europa. <https://drive.google.com/file/d/1r4PNEYBumN-6qSPtnu34qend0rLQfhyL/view>
- Lignon, Stéphanie & Namer Fiammetta. 2014. Les noms de procès en -ion : Quand le verbe appelle le verbe. In Florence Villoing, Sophie David & Sarah Leroy (éds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*. Nanterre : Presses Universitaires de Paris Ouest. [http://fiamm.free.fr/Publications/2014\\_HommagesFK\\_Lignon\\_Namer.pdf](http://fiamm.free.fr/Publications/2014_HommagesFK_Lignon_Namer.pdf)
- Marchand, Hans. 1960. *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*. Wiesbaden: Otto Harrassowitz.
- Marchand, Hans. 1969. *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*. München: C.H. Beck.
- Nagano, Akiko. 2007. Marchand's analysis of back-formation revisited. *Acta Linguistica Hungarica*, 54(1), 33-72.
- Oxford University Press (ed.). 2023. *Oxford English Dictionary [online]*. [www.oed.com](http://www.oed.com)
- Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin: Mouton de Gruyter.
- ATILF - CNRS & Université de Lorraine (éds.). 1994. *TLFi: Trésor de la langue Française informatisé*. <http://www.atilf.fr/tlfi>

---

# On Imaginary English Dvandvas in Relational Adjectives

Ryohei Naya

University of Tsukuba

Takashi Ishida

Hiroshima Shudo University

---

## 1 Typologically Unavailable, but Derivationally Available?

Dvandva compounds, a type of coordinated compound, have typological significance because they are widely observed in Asian languages, but not in European languages (Bauer (2008), Shimada (2013, 2016), among others). Thus, Japanese has the typical example of dvandvas expressing “a new unity made up of the whole of the two entities named” (Bauer (2008: 2)): *dan-jo* (male-female) ‘male and female.’ In contrast, its (Present-day) English counterpart is, as the translation shows, a phrase rather than a compound.

However, English has derivatives that apparently involve dvandvas. Such words can be observed in relational adjectives (RA), a kind of denominal adjective (e.g., *theatrical*, *historic*). Given that they have nominal bases, we encounter a paradoxical situation in the examples in (1) with combining forms composing neoclassical compounds (cited from the *OALD* and the *OED*): They appear to be derived from nominal dvandvas, which are supposed to be typologically unavailable in English.

- (1)
- |    |                  |  |
|----|------------------|--|
| a. | gastrointestinal | ‘of or related to the stomach and intestines’          |
| b. | dorsabdrominal   | ‘relating to the back and abdomen’                     |
| c. | oesophagogastric | ‘of or relating to the oesophagus and the stomach’     |
| d. | psychosomatic    | ‘involving or depending on both the mind and the body’ |

In (1a), *gastrointestinal* appears to contain as its base *\*gastrointestine*, which clearly has the typical reading of dvandvas ‘(the set of) the stomach and intestines,’ but this potential base is not a grammatical compound (Shimada (2023: 239)). Then, how can the RAs like those in (1) accommodate such an “imaginary” base, so to speak?

We aim to answer this question, drawing on Nagano’s (2013, 2015) analysis of RAs as prenominal variants of PPs, where P is a category-shifting functional category that turns an NP into an AP (Baker (2003)). If so, the RAs in (1) also have PPs as their underlying structures, where the nouns can be safely coordinated as in ordinary PPs (e.g., *in Europe and Asia*).

## 2 Framework: Nagano (2013, 2015)

The core idea of Nagano’s (2013, 2015) study is that RAs are morphological, realizational variants of PPs that appear in the environment of direct modification, where an attributive modifier is directly related to the head noun through base-generation. An important fact in this regard is that RAs can be semantically paraphrased as PPs, as in (2).

- (2)
- |    |                           |     |                               |
|----|---------------------------|-----|-------------------------------|
| a. | <i>presidential</i> plane | a’. | plane <i>of the president</i> |
| b. | <i>theatrical</i> dancer  | b’. | dancer <i>in the theater</i>  |

(Nagano (2013: 123; 2015: 6), with slight modifications)

Syntactically, this indicates that noun modification requires the modifier to be in the form of PP in the postnominal position and the form of RA in the prenominal position; PP cannot be a prenominal modifier as it stands (cf. *\*a* [<sub>pp</sub> *near* [<sub>DP</sub> *Boston*]] *residential area* (Escribano (2004: 2))).

Nagano (2013) then proposes that RAs are derived from the structure in (3a) through conflation (i.e., incorporation before lexical insertion). Specifically, N in (3a) is conflated into its head, P, forming the structure in (3b).



to form an RA. If so, it would be expected that the semantic subtypes of dvandvas attested in dvandva-rich languages could also be observed in English, albeit in the form of RAs. However, this is not the case. Particularly relevant here are what Bauer (2008) calls the co-synonymic and co-hyponymic types, which are exemplified in (8a, b), respectively. The co-synonymic dvandva consists of constituents in a synonymous relationship. The co-hyponymic type is a compound where each constituent denotes a subclass of the category named by the compound as a whole.

- (8) a. Co-synonymic    Lezgian    *kar-k'walax*    job work    'job, business'  
 b. Co-hyponymic    Punjabi    *bas-kaar*    bus-car    'vehicles'

(cited from Bauer (2008: 10, 9))

As dvandvas are not available in Present-day English, these subtypes are also systematically unobservable, and this situation holds true even in the form of RA.

First, the co-synonymic dvandvas composed of a combining form and its free form synonym would be something like *\*gastrostomachic* or *\*enterointestinal*, but these combinations are not easily acceptable.

Second, co-hyponymic dvandvas are also difficult to find in the RAs in question because, in most cases, as observed in (1), the coordinated expression simply refers to the union of the two sets named by the constituents, not exceeding it. One potential candidate for this kind of dvandva is *psychosomatic*, given the Japanese nominal dvandva *shin-shin* (mind-body), which can be used to express 'every fiber of one's being,' where *mind* and *body* can be understood as parts of one's existence. However, this meaning is not reflected in *psychosomatic*, and again, it simply denotes the sum of *mind* and *body*. In fact, *gastrointestinal tract* means the entire digestive tract, where *gastr-* (i.e., *stomach*) and *intestine* are both hyponyms of *digestive organ*. This is similar to the case of the co-hyponymic dvandva observed above, in which the coordinated hyponyms form their hypernym. This appears to be a potential challenge for our analysis. We assume that the RA formed in the proposed manner can undergo such a semantic extension (i.e., synecdoche), depending on its relationship with the noun to be modified. In fact, *gastrointestinal* is not always used to represent digestive organs; *gastrointestinal radiography* most likely refers to radiography of the stomach and intestines, not of the digestive tract as a whole. Thus, it is the modifier-head relationship that allows for semantic extension, arguing against (co-hyponymic) dvandva formation.

#### 4 Implications from the Lack of Neoclassical Dvandvas

In English, verbal compounds, as well as dvandvas, are typologically unattested in the sense that N-V compounds are not directly formed by combining two bases (e.g., *\*to truck-drive* (Ackema and Neeleman (2004))). Instead, they can be obtained by applying back-formation to (nominal or adjectival) synthetic compounds (e.g., *to air-condition<sub>v</sub>* < *air-conditioning<sub>N</sub>*). This raises the question of why this process is applicable to synthetic compounds but not to the RAs in (1), which would otherwise be a potentially rich source of neoclassical dvandvas in English. One answer is blocking by the phrasal competitor, as in *\*gastrointestine* vs. *stomach and intestines* (cf. *\*male-female* vs. *male and female*; see Nishimaki (2022) for a related discussion). Our analysis implies another possible factor behind the situation in which dvandvas are not back-formed from RAs. A crucial difference between synthetic compounds and RAs lies in how they are formed. Synthetic compounds are outputs of compounding, and if we take the view that compounding is lexeme-internal syntax (cf. Aronoff (1994: 16)), their formation, regardless of the exact process, is driven by syntax and arguably by semantics as well. On the other hand, RAs are the realization forms that the structure [P + N] is forced to take in the syntactic context of direct modification (see Section 2). In this sense, the formal

alternation from PP to RA is “closer to inflection” (Nagano (2013: 113)), although the resulting word has the status of a derivative. This difference may determine the applicability of back-formation; the outputs of syntactic context-triggered (or inflection-like) word-formation, but not those of syntax-/semantics-driven word-formation, are likely to resist undergoing back-formation (and possibly some types of word-formation processes). This situation is reminiscent of Myers’ Generalization that “no derivational suffixes may be added to a zero-derived word, just as no such suffix may be added to an inflected word” (Myers (1984: 66)). Our analysis, together with this generalization, leads us to examine the relationships among the relevant processes, which further deepens our understanding of how morphology works.

## References

- Ackema, Peter & Ad Neeleman. 2004. *Beyond Morphology: Interface Conditions on Word Formation*. Oxford: Oxford University Press.
- Aronoff, Mark. 1994. *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, MA: MIT Press.
- Baker, Mark C. 2003. *Lexical Categories: Verbs, Nouns, and Adjective*. Cambridge: Cambridge University Press.
- Bauer, Laurie. 2008. Dvandva. *Word Structure* 1. 65–86.
- Escribano, José Luis González. 2004. Head-Final Effects and the Nature of Modification. *Journal of Linguistics* 40. 1–43.
- Koshiishi, Tetsuya. 2011. *Collateral Adjectives and Related Issues*. Bern: Peter Lang.
- Myers, Scott. 1984. Zero-derivation and inflection. In Margaret Speas & Richard Sproat (eds.), *MIT Working Papers in Linguistics* 7. 53–69.
- Nagano, Akiko. 2013. Morphology of Direct Modification. *English Linguistics* 30(1). 111–150.
- Nagano, Akiko. 2015. Eigo-no Kankei-Keiyoshi (Relational Adjectives in English). In Tetsuo Nishihara & Shin-ichi Tanaka (eds.), *Gendai-no Keitairon to Onseigaku/On'inron-no Siten to Ronten* (Perspectives and Arguments of Present-day Morphology and Phonetics/Phonology), 2–20. Tokyo: Kaitakusha.
- Nagano, Akiko & Masaharu Shimada. 2014. Morphological Theory and Orthography: Kanji as a Representation of Lexemes. *Journal of Linguistics* 50. 323–364.
- Nishimaki, Kazuya. 2022. Coordinated Phrases as Dvandvas: A Competition-Theoretic Perspective. In Lotte Sommerer & Evelien Keizer (eds.), *English Noun Phrases from a Functional-Cognitive Perspective*, 395–427. Amsterdam/Philadelphia: John Benjamins.
- OALD: Oxford Advanced Learner’s Dictionary* <<https://www.oxfordlearnersdictionaries.com>>
- OED: Oxford English Dictionary Online* <<https://www.oed.com/>>
- Shimada, Masaharu. 2013. Coordinated Compounds: Comparison between English and Japanese. *SKASE Journal of Theoretical Linguistics* 10(1). 77–96.
- Shimada, Masaharu. 2016. Eigo-ni okeru Toifukugogo-no Seiki-nitsuite (The occurrence of Coordinated Compounds in English). In Yoshiki Ogawa, Akiko Nagano & Akira Kikuchi (eds.) *Kopasu-kara Wakaru Gengohenka/Hen’i-to Gengoriron* (The Linguistic Change / Variation as seen from Corpora, and Linguistic Theories), 307–323. Tokyo: Kaitakusha.
- Shimada, Masaharu. 2023. Dvandva Fukugogo-no Kozo-o Kangaeru (Considering the Structure of the Dvandva Compound). In Hiroshi Yonekura, Akiko Nagano & Masaharu Shimada (eds.), *Eigo-to Nihongo-niokeru Toifukugogo* (Coordinated Compounds in English and Japanese), 191–239. Tokyo: Kaitakusha.
- Watanabe, Akira. 2010. Eigo-no Koto-wa Eigo-dake Miteitemo Wakaranai—Keiyoshi-o Megutte (English from a Typological Viewpoint: The Case of Adjectives). Lecture delivered at Tsuda College.



---

# Morpho-semantics of the French diminutive suffix *-et(te)*

Adèle Hénot-Mortier  
Massachusetts Institute of Technology

---

## 1 Background

French assigns grammatical gender (Masculine or Feminine) to nominals and is endowed with a quite productive “diminutive” suffix *-et/-ette*.

- (1) a.  $\text{maison}_F \rightarrow \text{maisonnette}_F$                       b.  $\text{balcon}_M \rightarrow \text{balconnet}_M$   
    ‘house’  $\rightarrow$  ‘small (cute) house’                      ‘balcony’  $\rightarrow$  ‘small (cute) balcony’

Because M-bases are often affixed with the M-variant of the diminutive (*-et*) and F-bases with the F-variant (*-ette*), traditional grammars implicitly assumed that *-et* and *-ette* were allomorphs dependent on the gender features of the base, and were linked to the same diminutive semantics. Milner (1989) however observed that *-ette* may attach to M-bases and *-et* to F-bases – a phenomenon we dub **gender-mismatch** – leading to a looser semantic relationship between the base and the derived form.

- (2)  $\text{char}_M \rightarrow \text{charette}_F$                       (3) a.  $\text{boule}_F \rightarrow \text{boulet}_M$   
     $\text{char}_M \xrightarrow{*} \text{charet}_M$                       ‘ball’  $\rightarrow$  ‘cannonball’  
    ‘chariot’  $\rightarrow$  ‘cart’                      b.  $\text{boule}_F \rightarrow \text{boulette}_F$   
    ‘ball’  $\rightarrow$  ‘small ball’

These pairs would be unexpected if the suffix simply agreed in gender with the base: rather, it seems that in at least certain cases, the suffix introduces its own gender (a phenomenon documented in other languages, cf. Kramer 2015).

## 2 Contribution

In this work, we bring support to a refinement of Milner’s observation *via* a more systematic analysis of the French lexicon. More specifically, we argue that frequency differences between (i) *-et* and *-ette* suffixation (ii) M-to-F vs F-to-M gender-mismatches (iii) the number of “true” diminutives in the *-et* and *-ette* data (w/o a mismatch) can be explained if we assume that (1) *-ette* is ambiguous between an allomorph of the (non-purely diminutive) suffix *-et* and another very productive and purely diminutive suffix *-ette*; (2) gender-mismatching forms results from a root-level operation, unlike most gender-matching ones.

### 2.1 Data analysis.

From a list of French words (346,200 entries), we extracted and filtered nouns ending in *-et* and *-ette*. Filtering involved (1) finding the base from which the word is derived using online resources (Larousse online dictionary, Wiktionary) and introspection; (2) verifying that the base is a nominal. The dataset was supplemented by pairs generated *via* pure introspection (not all of them being documented in dictionaries) – for a total of 262 nouns in *-ette* and 146

nouns in *-et*. Further statistics are compiled in Tab. 1 below. In this table, the green, blue and red cells refer to gender-preserving suffixation, F-to-M mismatches and M-to-F mismatches respectively. The single numbers in parentheses in columns 2 and 3 correspond to the number of true diminutives, for each count. Finally, for bases with both a *-ette* and a *-et* form (column 4), the numbers in parentheses follow the format (# true *-ette* diminutives/ # true *-et* diminutives).

Three observations can be extracted from these lexicographic data. **The first observation is that *-ette* suffixation is around 1.8 times more frequent than *-et* suffixation.** Generating *-ette*-forms by introspection also appeared easier, suggesting that *-ette* is overall more productive than *-et*.

Derived → Base ↓	<i>-ette</i> only	<i>-et</i> only	Both	Total
Feminine	186 (138)	15 (5)	32 (23/7)	233
Masculine	34 (12)	89 (54)	10 (3/6)	133
<b>Total</b>	220	104	42	366

Table 1: Dataset statistics.

**The second observation is that the proportion of gender-mismatches is higher for M-bases (M-to-F mismatch) than F-bases (F-to-M mismatch):**  $\hat{P}[-et\text{-form}|F\text{-base}] = 47/233 \sim 20\% < \hat{P}[-ette\text{-form}|M\text{-base}] = 44/133 \sim 33\%$  ( $p = .006$ ). The amplitude of this discrepancy is approximately the same as the one recorded for *-et/-ette* forms in general ( $33/20 \sim 1.8$ ). It also seems that F-to-M mismatching forms are very likely to cooccur with a non-mismatching form derived from the same base ( $32/32+15 \sim 68\%$ ); while the opposite seems to hold for M-to-F forms (only  $10/10+34 \sim 22\%$  of them appear in “triplets”).

**The third and last observation is that 70% of the non gender-mismatching forms appear to have a true diminutive semantics; while only 30% of the mismatching forms do,** in line with Milner’s observation about the semantic effects of gender-mismatch. **However, a gender asymmetry arises in both “match” and “mismatch” cases:** non-mismatching F-forms in *-ette* are more likely to be diminutive than non-mismatching forms in *-et*:  $\hat{P}[DIM|F\text{-base-ette}] = 138+23/186+32 \sim 74\% > \hat{P}[DIM|M\text{-base-et}] = 54+6/89+10 \sim 60\%$  ( $p = .02$ ). The same pattern holds for mismatching forms, although non-significant, potentially due to small sample sizes:  $\hat{P}[DIM|M\text{-base-ette}] = 12+3/34+10 \sim 34\% > \hat{P}[DIM|F\text{-base-et}] = 5+7/15+32 \sim 26\%$ .

In brief, *-ette* appears more productive than *-et* and also more likely to lead to a diminutive semantics, and interestingly those two facts somewhat extend to mismatching forms (which were previously thought to be plain lexicalizations). We take this as evidence that *-ette* is (sometimes, at least) distinct from the allomorph of *-et*.

## 2.2 Formal analysis.

*Contra* previous accounts, we claim that ***-ette* is ambiguous between an allomorph of *-et* and a separate suffix *-ette*, which we assume is the pure French diminutive suffix DIM**, indicating relative smallness, cuteness, or affection towards the object. We take that *-et* has a looser semantics, which only involves a similarity with the base w.r.t. a salient feature, usually shape (so we write  $-et = \text{SHAPE}$  for brevity). This had been already noted by Milner (1989) and Delhay (1999), but mostly for gender-mismatch cases. Yet, pairs like those in (4) and (5) exemplify the same kind of loose semantic relationship in matching-gender cases, *for both genders* – in line with our ambiguity hypothesis. *-et* being the realization of SHAPE and *-ette* being that of either SHAPE + AGREE or DIM also explains why *-ette* is more frequent than *-et* across the board, and more likely to yield a diminutive semantics.

- (4) a.  $oeil_M \rightarrow oeillet_M$   
       ‘eye’ → ‘eyelet’
- b.  $arc_M \rightarrow archet_M$   
       ‘bow (archery)’ → ‘bow (music)’

- (5) a. barre<sub>F</sub> → barrette<sub>F</sub>      b. coquille<sub>F</sub> → coquille<sub>F</sub>  
       ‘bar (construction)’ → ‘hair-clip’      ‘shell’ → ‘elbow pasta’

Our second claim, which builds on the Lexical Decomposition hypothesis (Marantz 1997, 2001; Arad, 2003, 2005), is that **gender-mismatching forms result from a merger of the DIM/SHAPE suffix at the root-level, unlike gender-matching forms, whose suffix is merged above the nominalizing-head *n* (which we assume hosts gender features)**. In the “mismatch” case, the suffix *is* the categorizing head and therefore imposes its own gender on the root; in the “match” case, the suffix follows (and agrees with) the gender already introduced by *n*. Following Arad (2003), we also argue that the root-level derivation generating gender-mismatching forms introduces additional semantic noise, due to the uncategorized root having an underspecified meaning. This explains why gender-mismatching forms are less likely to be diminutive, *while still exhibiting a gender-related asymmetry* (M-to-F vs F-to-M). In particular, we predict M-to-F forms in *-ette* to exhibit a diminutive semantics (contributed by *-ette*, which is unambiguously DIM in that case), but not on the “right” entity (due to root-underspecification). This might be the case for the pairs in (6) below.

- (6) a. cigare<sub>M</sub> → cigarette<sub>F</sub>      b. disque<sub>M</sub> → disquette<sub>F</sub>  
       ‘cigar’ → ‘cigarette’      ‘CD/hard disk’ → ‘floppy disk’

### 3 Conclusion, and a remaining puzzle

We argued that the difference in productivity and transparency between *-ette* and *-et* was due to *-ette* being ambiguous between an allomorph of *-et* (not purely diminutive) and DIM. We showed the discrepancy was modulated by gender-mismatches, which we argued were the result of root-level derivation and therefore linked to extra semantic noise. The full set of predictions is summarized in Tab. 2.

Crucially, our account provided a morphosyntactic explanation as to why gender-mismatches correlate with some form of *semantic* mismatch. Previous accounts positing lexicalization did not really address this issue.

Base	Suffix	Level	Form	Semantics
M	SHAPE	1/2	<i>-et</i>	loose on (noisy) root
	DIM	1	<i>-ette</i>	dim. on noisy root
F	SHAPE	1	<i>-et</i>	loose on noisy root
	SHAPE + AGR	2	<i>-ette</i>	loose on exact root
	DIM	1/2		dim. on (noisy) root

Table 2: Summary of the predictions. ‘1’ = root-level derivation; ‘2’ = above *n*

A remaining puzzle is the following: why are 60/99 M-forms in *-et* diminutive, given that we predict the more general SHAPE relationship to hold in that case? We think this may be due to some form of morphological reanalysis targeting a specific subset of the *-et*-forms. Indeed, a DIM-meaning is more likely to arise for bases ending in *in/on/eau* (38/41), which already have a fossilized diminutive flavor:<sup>1</sup> Such endings were also the preferred targets for applying *-et* productively. This suggests that they were perhaps re-analyzed as proper morphemes (contributing the DIM semantics) by the action of *-et* suffixation.

<sup>1</sup>We use this denomination because most of the nominals from the dataset with such endings (e.g. *cochon*, ‘pig’, *champignon*, ‘mushroom’) were morphologically simplex; yet, the same endings are common in proper names (*Antoine* → *Antonin*; *Marie* → *Marion*; *Boucher* → *Bouchereau*...) and appear consistently diminutive.

## References

- Arad, Maya. 2003. Locality constraints on the interpretations of roots. *Natural Language and Linguistic Theory* 21(4). 737–778. doi:10.1023/a:1025533719905.
- Arad, Maya. 2005. *Roots and patterns: Hebrew morpho-syntax*. Dordrecht: Springer.
- Aronoff, M. 1976. *Word formation in generative grammar*.
- Booij, Geert. 2007. *The Grammar of Words: An Introduction to Linguistic Morphology*. Oxford University Press. doi:10.1093/acprof:oso/9780199226245.001.0001. <https://doi.org/10.1093/acprof:oso/9780199226245.001.0001>.
- Deal, Amy Rose. 2016. Plural exponence in the nez perce DP: a DM analysis. *Morphology* 26(3-4). 313–339. doi:10.1007/s11525-015-9277-9. <https://doi.org/10.1007/s11525-015-9277-9>.
- Delhay, Corinne. 1999. "diminutifs" et niveaux de catégorisation. *Faits de langues* 7(14). 79–87. doi:10.3406/flang.1999.1268. <https://doi.org/10.3406/flang.1999.1268>.
- Dressler, Wolfgang U. & Lavinia Merlini Barbaresi. 2001. Morphopragmatics of diminutives and augmentatives. In *Perspectives on semantics, pragmatics, and discourse*, 43–58. John Benjamins Publishing Company. doi:10.1075/pbns.90.07dre. <https://doi.org/10.1075/pbns.90.07dre>.
- Ferrari, F. 2005. *A syntactic analysis of the nominal systems of italian and luganda: how nouns can be formed in the syntax*: New York University dissertation.
- Gougenheim, G. 1946. Les féminins diminutifs en français moderne. *Modern Language Notes* 61(6). 416. doi:10.2307/2908930. <https://doi.org/10.2307/2908930>.
- Jurafsky, Daniel. 1996. Universal tendencies in the semantics of the diminutive. *Language* 72(3). 533. doi:10.2307/416278. <https://doi.org/10.2307/416278>.
- Kornexl, Lucia. 2008. Women and other 'small things': -ette as a feminine marker. In *English historical linguistics 2006*, 241–257. John Benjamins Publishing Company. doi:10.1075/cilt.296.16kor. <https://doi.org/10.1075/cilt.296.16kor>.
- Kramer, Ruth. 2015. *The morphosyntax of gender* Oxford Studies in Theoretical Linguistics 58. Oxford: Oxford University Press.
- Kramer, Ruth. 2016. The location of gender features in the syntax. *Language and Linguistics Compass* 10(11). 661–677. doi:10.1111/lnc3.12226. <https://doi.org/10.1111/lnc3.12226>.
- Lyster, Roy. 2006. Predictability in french gender attribution: A corpus analysis. *Journal of French Language Studies* 16(1). 69–92. doi:10.1017/S0959269506002304.
- Marantz, Alec. 1997. No escape from syntax. *University of Pennsylvania Working Papers in Linguistics* 4. 201–225.
- Marantz, Alec. 2001. *Words*.
- Milner, Jean-Claude. 1989. Genre et dimension dans les diminutifs français. *Linx* 21(1). 191–201. doi:10.3406/linx.1989.1141. <https://doi.org/10.3406/linx.1989.1141>.
- Nelson, Don. 2005. French gender assignment revisited. *Word* 56(1). 19–38.
- Plénat, Marc & Michel Roché. 2001. Prosodic constraints on suffixation in French. In *Proceedings of the Third Mediterranean Morphology Meeting*.
- Roché, Michel. 1992. Le masculin est-il plus productif que le féminin ? *Langue française* 96. 113–124.
- Schneider, Klaus P. 2003. *Diminutives in english*. Berlin, Boston: Max Niemeyer Verlag. doi:10.1515/9783110929553. <https://doi.org/10.1515/9783110929553>.

---

# Affix rivalry in French demonyms: an experimental approach

Marie Huygevelde Ridvan Kayirici Olivier Bonami Barbara Hemforth

Université Paris Cité, Laboratoire de linguistique formelle, CNRS

---

This study reports on two acceptability experiments on the influence of base phonology on speakers' preferences in the choice of a suffix forming a demonym (e.g. *nancéen*, 'from Nancy') from a toponym (e.g. *Nancy*). The results are broadly consistent with previous findings by Thuilier et al.'s (2023), but bring out interesting contrasts between tendencies in the established lexicon and speakers' preferences in novel formations.

## 1 Motivation

Ever since Aronoff (1976) coined the term, situations of rivalry, where multiple word formation processes are available to convey the same meaning, have been a major focus of attention for descriptive and theoretical morphology. Much progress in this area has been made possible by the systematic exploration of large lexical databases (e.g. Plag 1999; Lindsay & Aronoff 2013) and the application to these of statistical modelling (Baayen et al., 2013) and computational simulations (Arndt-Lappe, 2014).

One important limitation of this line of work is the inherent heterogeneity of the data found in the established lexicon, which contains words coined over centuries by speakers whose linguistic experience may have differed significantly—if anything because each new coinage may influence later formations. While some authors attempt to alleviate this problem by focusing on recent formations (Plag, 1999) or explicitly taking into account diachronic variation (Lindsay & Aronoff, 2013; Arndt-Lappe, 2014), a more direct (but less frequent) approach is to conduct behavioral experiments probing the preferences of speakers facing the task of producing, interpreting or judging rival formations (Anshen & Aronoff, 1981; Makarova, 2016).

This abstract reports on such behavioral experiments dedicated to the formation of demonyms in French. Demonyms are a particularly promising testbed for such a study: on the one hand, the semantic relationship between a toponym and its demonym is much more stable than what is found for other instances of word formation; hence we need not worry about sibilant decisions as to whether a particular pair of words instantiates the relevant morphosemantic contrast. On the other hand, rivalry is very prevalent, with at least four highly productive suffixes in French: *-ais* (see *Marseillais* from *Marseille*), *-ois* (*Lillois*, from *Lille*), *-éen* (*Nancéen*, from *Nancy*) and *-ien* (*Parisien*, from *Paris*). It is also very well documented, in no small part thanks to Thuilier et al.'s (2023) recent thorough study of more than 2,000 established toponyms. Thuilier et al.'s study serves as the starting point for the present research: our goal is to assess the extent to which speaker preferences in an experimental setting track the tendencies observed by Thuilier et al. in the established lexicon. We focus more specifically on the influence of phonological properties of the base on the choice of a suffix when speakers are faced with a novel, unknown toponymic base.

## 2 Methods

To explore the impact of the phonological makeup of toponyms in demonym formations, we ran two experiments investigating preferential choices with disyllabic toponymic bases with a final consonant (Experiment 1) or nasal vowel (Experiment 2). We limited our attention to the

four most productive suffixes *-ais*, *-ois*, *-éen* and *-ien* that have been highlighted in the literature (Eggert, 2002; Plénat, 2008; Thuilier et al., 2023)).

For the first experiment, we created 24 toponyms with a final segment representative of each of the 5 categories singled out by Thuilier et al.<sup>1</sup>. Our toponyms ended with the bilabial plosive /p/ (Nabope), the palatal approximant /j/ (Naboje), the alveolar fricative /s/ (Nabosse), the post-alveolar fricative /ʃ/ (Naboche), and the nasal bilabial /m/ (Nabome) (standing for the ‘other segment’ category in Thuilier et al.’s study).

For the second experiment, we explored the impact of vowel backness on nasal last segments based on tongue positioning. We created 15 items with proper names manipulating the backness of both nuclei in a disyllabic word ending in a nasal vowel. Thus, our vocalic feature pairs consisted of /a/ and /o/ for the first position, /ẽ/ and /õ/ for the second position. We made use of the methodology proposed by Lohmann (2017) to control for the backness score of the base toponyms, we had thus vocalic nasal feature pairs with low /a/ /ẽ/, high /o/ /õ/, and medium /a/ /õ/ and /o/ /ẽ/ backness scores.

Our experiments took place online, using a local installation at LLF of Alex Drummond’s IBEX software. Each experimental session started with written instructions, a short anonymized questionnaire, and a brief practice session.

During each trial, a base toponym was presented to the participants and they were asked to choose between four possible demonyms formed using each of the four main suffixes with a given toponym. Participants saw all conditions but each item in only one condition following a Latin square design. In total, an experimental session consisted of 24 trials with consonant-final toponyms, 15 trials with nasal vowel-final toponyms, and 11 trials of fillers based on real-world toponyms (i.e. Marseille). Items and fillers were presented in random order.

We recruited 71 participants from online academic and social networks. Two bilingual participants were excluded from our analyses (n = 69, mean age = 41). All participants took the experiment voluntarily with no compensation and gave their consent for the usage of their anonymized data for scientific purposes.

### 3 Results

All data were analyzed with Generalized linear mixed models (binomial family), in R (R version 4.1.1, glmer, lme4 version 1.1-31), with random intercepts for participants and item.<sup>2</sup>

Across the two experiments, the dependent variable was the choice between four categorically distributed nominal values that are constituted by the demonym suffixes *-ais*, *-ois*, *-éen* and *-ien*. The phonological final segment constraints constituted the independent variable (5 conditions) for the consonantic demonym formations, and the backness feature as well as the syllabic position were the independent variables for the vocalic nasal demonym formation.

Figure 1 shows that all suffixes are used in all conditions but with clearly varying frequencies. We observed a preference for the suffix *-ien* for the bilabial plosive, alveolar fricative and palatal approximant conditions, while the post-alveolar fricative condition was more associated with the suffix *-ois*. For the consonantic nasal bilabial condition, we observed a tendency towards the suffixes *-éen* and *-ois* in a similarly weighted manner accounting for more than half of the preferences combined.

For statistical comparisons, the palatal fricative /ʃ/ served as the reference category. The suffix *-ois* was chosen significantly more often for the palatal fricative than for the bilabial plosive ( $p < .001$ ) as well as the palatal approximant ( $p < .001$ ), followed by /m/ ( $p < .04$ ) and

<sup>1</sup>One item of each experiment had to be excluded for technical reasons

<sup>2</sup>Random slopes were not included because of convergence failures.

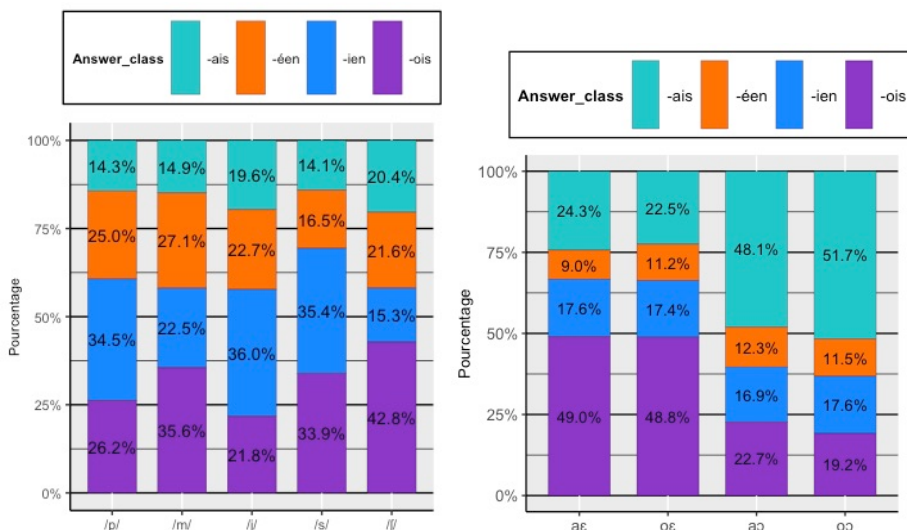


Figure 1: Suffix choices for Experiments 1 and 2

/s/ ( $p < .01$ ). The suffix *-ien* was chosen least frequently for the palatal fricative /ʃ/ compared to all other phonological conditions: /s/ ( $p < .001$ ), /j/ ( $p < .001$ ), /p/ ( $p < .001$ ), /m/ ( $p < .02$ ). The suffix *-ais* was chosen more frequently in the palatal fricative than in the alveolar fricative ( $p < .04$ ) and bilabial plosive ( $p < .05$ ) conditions. For the suffix *-éen*, we did not observe any significant differences across the five conditions.

The second experiment showed significant preferences for the suffix *-ais* for the toponyms with a back last segment ( $p < .001$ ), and for *-ois* for the toponyms with a front last segment ( $p < .001$ ). We did not find any significant differences concerning the first syllable position or the backness score. We did not observe any significant effects for choices of *-éen* and *-ien*.

## 4 Discussion

Thuilier et al.'s study established an overall preference for the suffix *-ois* after consonants. In general, we observed similar results aligning with this tendency, in particular for the alveolar and post-alveolar fricative conditions.

For the alveolar fricative condition, Thuilier et al. documented 49.6% of use for *-ien*, and 37.6% for *-ois*. We found a similar tendency, with 35.4% for *-ien*, and 33.9% for *-ois*. These results might be indicative of a phonological motivation caused by the alveolar fricative feature, which is also captured in the corpora. For the post-alveolar fricative condition, *-ois* was chosen 42.8% of the time by the participants. While this made up most of their choices, the preference is less sharp in our experiment compared to its prevalence in the established lexicon where it reaches 69.5% in the "other fricative" condition.

For the bases ending with a bilabial plosive or a palatal approximant, we found contrasting results between data from experiments and corpora. Whereas 47.2% of *-ois* was found after plosives in the corpus, it accounted for only 26.2% of the choices in our experiment. Likewise, while *-ois* was the preferred option in the established lexicon after an approximant with 57.2% (vs. 20.5% for *-ien*), in our experiment *-ien* was preferred at 36% (only 21.8% for *-ois*).

In the nasal bilabial condition, we found both similarities and discrepancies with what Thuilier et al. observed in the established lexicon. Our participants showed the same preference for *-ois* documented by Thuilier et al. for the distribution of the suffixes with respect to nasal segments, although less strongly so (48.4% vs. 35.6%). When we consider the bilabial

nasal condition from 'other segment' perspective, where *-ois* rate drops to 27.9% (after suffixes *-ien* and *-ais* 32.3% and 28.2% respectively) in the established lexicon, we observe a slight discrepancy. This might be due to the fact that the 'other' lumps together a diverse set of segments in their study, while in our experiments, the consonant is specified as /m/. However, our results are surprising in another dimension. In the affix rivalry literature, it has been argued that dissimilative tendencies disfavor *-éen* and *-ien* after a nasal segment, be it a vowel or consonant. We found no evidence for this, as *-éen* and *-ien* made up half of the choices after /m/.

Thuilier et al.'s data suggest that the nasality of the vowel plays an important role in the choice of a suffix. Our findings suggest that demonym formations with nasal vowels might be strongly motivated by phonological constraints. Keeping in mind that our vocalic conditions involve also nasality, findings coming from our second experiment align with suggestions from the literature on vocalic qualities as indicated by Thuilier et al., and highlighted by Eggert, Plénat, and Roché & Plénat (2016): we see that the suffix *-ais* is disfavored after bases with last segment front vowel, whereas the suffix *-ois* disfavors bases with last segment back vowel.

One main conclusion of this study is that phonological constraints seem to play a distinct role for affix rivalry dynamics regarding French demonym formations. The discrepancies between observations from our controlled experiments and the results from the Thuilier et al. study might be due to factors beyond phonological motivations that may play a role for "real" demonyms but also to specifics of the experimental task.

## References

- Anshen, Frank & Mark Aronoff. 1981. Morphological productivity and phonological transparency. *Canadian Journal of Linguistics* 26. 63–72.
- Arndt-Lappe, Sabine. 2014. Analogy in suffix rivalry: the case of English *-ity* and *-ness*. *English Language and Linguistics* 18. 497–548.
- Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge: MIT Press.
- Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nessel. 2013. Making choices in Russian: pros and cons of statistical methods for rival forms. *Russian Linguistics* 37(3). 253–291.
- Eggert, Elmar. 2002. *La dérivation toponymes-gentilés en français : mise en évidence des régularités utilisables dans le cadre d'un traitement automatique*: Tours, Phd thesis.
- Lindsay, Mark & Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. In Nabil Hathout, Fabio Montermini & Jesse Tseng (eds.), *Morphology in toulouse: Selected proceedings of décembrettes 7*, 133–153. Munich: Lincom Europa.
- Lohmann, Arne. 2017. Phonological properties of word classes and directionality in conversion. *Word Structure* 10(2). 204–234.
- Makarova, Anastasia. 2016. Variation in Russian verbal prefixes and psycholinguistic experiments. In Tanja Anstatt, Anja Gattnar & Christina Clasmeier (eds.), *Slavic languages in psycholinguistics*, 113–133. Tübingen: Narr Francke Attempto Verlag.
- Plag, Ingo. 1999. *Morphological productivity*. Berlin: Mouton de Gruyter.
- Plénat, Marc. 2008. Quelques considérations sur la formation des gentilés. In *La raison morphologique*, 155–174. John Benjamins.
- Roché, Michel & Marc Plénat. 2016. De l'harmonie dans la construction des mots français. *SHS Web of Conferences* 27. 08003.
- Thuilier, Juliette, Delphine Tribout & Marine Wauquier. 2023. Affixal rivalry in French demonym formation: The role of linguistic and non-linguistic parameters. *Word Structure* 16(1).



---

# Creativity in name-based word formation: Evidence from the experimental study of personal name blends

Milena Belosevic

Bielefeld University

---

## 1 Introduction

Word formation units with proper names as constituents are often defined as creative, playful (cf. Beliaeva 2019) or extra-grammatical (cf. Mattiello 2013). Given the special formal and semantic properties of names compared to lexical units (cf. Anderson 2007), it remains unclear how name-specific aspects interact with speakers' linguistic experience in the production and processing of name-based word formation and whether this interaction contributes to their creativity.

The present paper aims to account for these issues by looking at the creativity of personal name blends (henceforth PN blends, e. g., *Brangelina* from *Brad* and *Angelina*). PN blends are a rather under-researched phenomenon that has not been systematically investigated on the basis of experimentally elicited data. Whereas recent studies on lexical blends (e.g., *brunch* from breakfast and lunch) provide evidence for their structural predictability regarding, for instance, the blend structure and the order of constituents, and put the creativity of lexical blending into question (cf. e.g., Gries 2012), the creativity of PN blends has not been investigated systematically.

Starting from the hypothesis that PN blends bear formal and semantic similarities with lexical blends and binomials such as *Romeo and Julia* (cf. Filatkina et al. 2019), PN blends are regarded as creative if they deviate from the properties of lexical blends and binomials. In particular, the creativity of PN blends is operationalized in terms of a deviation from the conventional constituent order in lexical blends and binomials. In this regard, the hypotheses about the regularities regarding the order of constituents in lexical blends proposed by Kelly (1998) are tested on experimentally elicited PN blends. As it will be shown, the results indicate that PN blends are rather not creative since language users tend to conform to the order of constituents underlying binomials and lexical blends.

## 2 Name blending

In this paper, a schema-based approach to blending (cf. Kemmer 2003) is adopted according to which PN blends emerge from cognitively entrenched patterns of experience with the order of constituents in lexical blends, binomials, but also in very frequent PN blends such as *Brangelina* or *Bennifer*. The paper is concerned with ascriptive PN blends in which both names are equally important semantically (cf. *Brangelina*). They do not bear a modifier-head structure and are similar to coordinative compounds (cf. Kotowski et al. 2021).

To test whether name constituents in PN blends deviate from the conventional order of constituents and can be defined as creative, the paper draws on the study of the constituent order in lexical blends provided by Kelly (1998). Kelly's approach is suitable for testing the creativity of PN blends for two reasons: The definition of lexical blends as contractions of conjunctive phrases (cf. Kelly 1998) allows for a comparison between lexical blends and ascriptive PN blends. Furthermore, Kelly's study is explicitly concerned with the interaction between linguistic and non-linguistic factors in ordering the blend constituents and therefore allows it to account for the above-mentioned issues regarding the interaction of name-specific and linguistic factors in the formation of creative name-based units. His study focuses on three

factors: the syllabic length of constituents, the frequency of constituents, and their prototypicality. The following two hypotheses regarding the conventional order of lexemes in lexical blends are proposed: 1) Shorter and more frequent constituents of lexical blends occupy the first position. 2) More prototypical, more frequent, and shorter constituents occupy the first position in lexical blends. In this paper, prototypicality is operationalized using the concept of familiarity with name constituents in that familiarity with names is considered a complex phenomenon comprising both frequency and prototypicality (cf. Zimmer 2018 for a similar operationalization). Apart from the syllabic length of name constituents, the (biological) gender of name bearers as a name-specific factor has been included in the analysis.

### 3 Production experiment

Since to the author's knowledge, experimental studies on PN blends in German do not exist and corpus data do not allow for controlling single factors, a production experiment was conducted to test the following hypotheses regarding the order of name constituents in PN blends:

- Hypothesis 1: Familiar, male and shorter names are preferred in the first position over unfamiliar female and longer names.
- Hypothesis 2: Familiar male names occupy the first position compared to unfamiliar female first names (given the same syllabic length of both names).
- Hypothesis 3: Familiar and shorter first names occupy the first position compared to unfamiliar and longer name constituents (given the same gender of name constituents).
- Hypothesis 4: Male and shorter names occupy the first position (if both constituents are familiar or unfamiliar).

Note that the “male first”- hypothesis has been derived from the studies on constituent order in binomials (cf. Cooper & Ross 1975: 65). PN blends that correspond to hypotheses 1 to 4 are regarded as conventional/rather not creative in terms of the constituent order. Otherwise, they are regarded as creative.

#### 3.1 Stimuli, participants and procedure

The stimuli comprised 56 name pairs (e.g., *Stefan and Renate*, *Lisa and Salihe*, *Torsten and Gratian*) formed from 16 male and 16 female first names. The names were controlled for the following factors: gender (male and female), syllabic length (bi- and three-syllabic names), and familiarity with name constituents (familiar or unfamiliar). In both groups, eight first names were disyllabic (four male and four female) and eight were three-syllabic (four male and four female). Familiar and unfamiliar names were selected from the list of the 50 most familiar and most unfamiliar first names in Germany. The list is a result of a rating experiment conducted as a part of a longitudinal study “The image of names”<sup>1</sup>.

Since female and male first names bear different formal properties in German, the stress position, length (in terms of the number of syllables, distribution of vowels and consonants, and final position) were controlled using the gender index for first names in German (cf. Nübling 2017). Note that, in German, only first names indicate biological gender. The syllabic length of names (bi- and three-syllabic) is based on the fact that the average length of prototypical German male first names is 1.92 compared to the average length of female names (2.54 syllables, cf. Nübling 2017: 107). The items were grouped into four conditions so that in each condition the interaction between two variables was investigated and one variable was controlled: 1) familiar

<sup>1</sup> Cf.: [https://www.onomastik.com/Vornamen-Lexikon/feature\\_ranking.php?feature=1&gender=](https://www.onomastik.com/Vornamen-Lexikon/feature_ranking.php?feature=1&gender=)

male disyllabic name + unfamiliar female three-syllabic name, 2) familiar male name + unfamiliar female name (same length), 3) disyllabic familiar name + three-syllabic unfamiliar name (same gender), 4) disyllabic male name + three-syllabic female name (same familiarity).

45 students (73 % native speakers of German and 27 % bilinguals, 80 % female and 20 % male) enrolled in the Linguistic programme (average age 24.1 years, SD = 3.6) participated in the experiment. As reported in a post-questionnaire, the majority of participants (80 %) had experience with lexical and name blends. Each participant was exposed to all 56 items in all four conditions. To minimize the effect of the order of presentation, the participants were divided into five groups of nine. In each group, the order of conditions and items varied. The participants were asked to build a new name from name pairs presented in the context assumed to be typical of PN blends, namely a romantic relationship, by shortening both names or only one of them without adding new letters. Although the definition of blends provided in section 2 was a part of the instructions given to participants, the term *blend* was not explicitly mentioned in the task description. The task was performed as a web experiment without time pressure.

### 3.2 Results

The production study yielded 2752 tokens (31 % hapaxes). Blends that did not comprise the beginning of the first and the ending of the second name (cf. Gries 2012: 146 for lexical blends), such as *Nihanna* from *Nina* and *Johannes*, or so-called clipping compounds (*Chrisle* from *Christofer* and *Lena*) were excluded from the analysis so that 2193 tokens (38 % hapaxes) were investigated. The blends were manually annotated for the order of constituents, the gender of names (male or female), their length (di- or three-syllabic names), and familiarity (familiar or unfamiliar). Afterwards, each condition was analysed using Pearson's chi-square test for goodness of fit to measure whether the difference between the observed distribution of name order and a random distribution is statistically significant.

First names that are familiar, male and disyllabic (e.g., *Martin*) are placed in the first position in 60 % over unfamiliar, female and three-syllabic names (e.g., *Salihe*) that occupy the second position (cf. hypothesis 1). The distribution of conventional and nonconventional variants is statistically significant ( $\chi^2 = 9.8$ ,  $p = 0.001$ ,  $df = 1$ ).

Regarding the second hypothesis, PN blends with a familiar and male name in the first position (e.g., *Christide* < *Christofer* and *Hamide*) occur in 50 % compared to the unconventional variants with an unfamiliar female name of the same length in the first position, e. g., *Hamofer*). However, this distribution is not statistically significant ( $\chi^2 = 0.01$ ,  $p = 0.89$ ,  $df = 1$ ) and the preference for the conventional form does not account for cases where both names are disyllabic.

The preference for the more familiar and shorter constituent has been confirmed for the combinations of two names of the same gender but different familiarity and length (e.g., *Torstian* vs. *Grasten* from *Torsten* and *Gratian*, hypothesis 3). The conventional order of name constituents occurs in 57 %, a familiar and disyllabic name occupies the first position more frequently than an unfamiliar three-syllabic name. Although the distribution of a conventional and nonconventional constituent order is statistically significant ( $\chi^2 = 10.5$ ,  $p = 0.001$ ,  $df = 1$ ), it is only true if both names are male.

When familiarity is controlled, a male shorter name occupies the first position in 60 % compared to a female longer name constituent (e.g., *Stenate* wins over *Refan* < *Stefan* and *Renate*), confirming the fourth hypothesis. This distribution is statistically significant ( $\chi^2 = 7.08$ ,  $p = 0.007$ ,  $df = 1$ ).

To sum up, the production study yielded that participants rarely deviate from the conventional constituent order in binomials and lexical blends so that PN blends cannot be regarded as creative in the sense of the definition provided in section 1.

## 4 Conclusions and outlook

This paper presented the results of a production study investigating the creativity of name-based word formation in terms of speakers' preference for a conventional or nonconventional order of name constituents in PN blends. The study of the interaction between linguistic factors (syllabic length of names) and non-linguistic (name-specific) factors (familiarity with names and the gender of name bearers) based on hypotheses about the conventional order of lexemes in lexical blends proposed by Kelly (1998) did not yield evidence for the status of PN blends as creative word-formation units from the perspective of the preferred constituent order. Furthermore, it can be concluded that not only the constituent order but also the blend structure are similar to the formal properties of lexical blends (ca. 80 % bearing the structure typical of lexical blends, i. e., the beginning of the first and the end of the second constituent). Finally, evidence for the interaction between name-specific aspects and speakers' linguistic knowledge is evident from the fact that the conventional order is constrained by name-specific properties, such as the gender of name bearers (cf. hypothesis 3). Future studies should address the role of extralinguistic factors related to the properties of language users, such as age and linguistic experience with blending. Given that Kelly's study included a limited number of variables, the interaction between further linguistic factors (e.g., the preference for particular switch points and transparency grades in combination with contextual factors) and extralinguistic aspects should be investigated to gain a more complete insight into the mechanisms underlying the formation of PN blends.

## References

- Anderson, John M. 2007. *The grammar of names*: Oxford, Oxford University Press.
- Beliaeva, Natalia (2019): Blending creativity and productivity: on the issue of delimiting the boundaries of blends as a type of word formation. In: *Lexis* 14.
- Cooper, William/John R. Ross. 1975. *World Order*. In: Robin E. Grossman et al. (ed.), *Papers from the Parasession on Functionalism*, Chicago, 63–111.
- Filatkina, Natalia et al. (2019) : *Mercron, Robbery und Donary Clump*. Die sog. Namenkreuzungen als bewegte Namen. Paper presented at the International meeting of the German society for name studies, Münster 2019.
- Gries, Stefan. 2012. Quantitative corpus data on blend formation: Psycho- and cognitive-linguistic perspectives. In: Renner, Vincent et al. (ed.): *Cross-Disciplinary Perspectives on Lexical Blending*. Berlin, 145–168.
- Kelly, Michael. 1998. To “brunch” or to “brench”: some aspects of blend structure. *Linguistics* 36, 579–590.
- Kemmer, Susanne. 2003. Schemas and lexical blends. In: Radden, Günter/Cuyckens, H. (Hrsg.): *Motivation in language. Studies in honor of Günter Radden*. Amsterdam, 69–97.
- Kotowski, Sven et al. 2021. The semantics of personal name blends in German and English. [https://www.anglistik3.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische\\_Fakultaet/Anglistik\\_und\\_Amerikanistik/Ang3\\_Linguistics/Dateien/Detailseiten/Kotowski/Kotowski\\_et\\_al\\_blends.pdf](https://www.anglistik3.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische_Fakultaet/Anglistik_und_Amerikanistik/Ang3_Linguistics/Dateien/Detailseiten/Kotowski/Kotowski_et_al_blends.pdf)
- Mattiello, Elisa. 2013. *Extra-Grammatical Morphology in English: Abbreviations, Blends, Reduplicatives, and Related Phenomena*. Berlin.
- Nübling, Damaris. 2017. Beziehung überschreibt Geschlecht. Zum Genderindex von Ruf- und Kosenamen. In: Linke, Angelika/Schröter, Juliane (ed.): *Sprache und Beziehung*. Berlin, 99–118.
- Zimmer, Christian. 2018. *Die Markierung des Genitiv(s) im Deutschen: Empirie und theoretische Implikationen von morphologischer Variation*. Berlin: de Gruyter.

---

# Baseless derivation: the behavioural reality of derivational paradigms

Maria Copot      Olivier Bonami  
Université Paris Cité    Université Paris Cité

---

## 1 Background

The historically dominant conceptualisation of the structure of derivational families is the ROOTED TREE, where each lexeme is either at the root of a derivational tree or has a unique parent. In this view, illustrated in Figure 1, lexemes are linked by monodirectional relationships, and there is a always single path relating two lexemes.

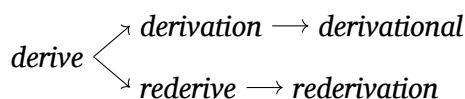


Figure 1: A rooted tree representation of part of the derivational family of *derive*.

This view incurs a number of theoretical and descriptive issues. The rooted tree view forces one to choose a single base for each derived lexeme, which is problematic, for instance, when a derived lexeme has multiple potential bases (does *asymmetrical* come from *asymmetric* or from *symmetrical*?), or when a lexeme’s semantics do not come from its formal base, but from another member of the derivational family (Hathout & Namer, 2014): the formal process  $X \sim Xics$  generally signifies a relationship between an object and the discipline that studies it (*graph*~*graphics*, *gene*~*genetics*), but the story is more complicated for the pair *language*~*linguistics* - *linguistics* has taken its meaning from *language*, but its base form from *linguist*, since *linguistics* is not the study of linguists. Moreover, the rooted tree’s requirement that relationships be monodirectional is ill-suited to capture cases of backformation, where a formally simpler lexeme can be shown to have been derived from a formally more complex one) or cross-formation (Becker, 1993), where two derived lexemes are more predictive of each other’s form and meaning than their common base is, or when the base is absent).

An alternative to the rooted tree view of derivational families is a paradigmatic approach (Robins, 1959; Becker, 1993; Bochner, 1993; Bauer, 1997; Štekauer, 2014; Bonami & Strnadová, 2019; Hathout & Namer, 2022). Seeing derivational families as paradigmatic involves foregrounding the multiple, bidirectional relationships that exist between related lexemes, positing no single lexeme as more basic than, or logically previous to, any other.

## 2 Motivation

The two views of morphology make different predictions about which relationships between word forms are accessible to speakers. A paradigmatic approach supposes that speakers keep track of all bidirectional relationships between word forms. The rooted tree view of word formation is traditionally associated with post-bloomfieldian morphemic approach to morphology, but such approaches haven’t traditionally engaged in making predictions about the cognitive reality of relationships of predictability within words (though see Jun & Albright 2016; Cotterell et al. 2019 for related examples in inflection). Word-based approaches to word formation

such as Aronoff (1976) and Stump (2019) also suggest that the canonical situation is for words of a language to be organised in a rooted tree structure, a view that carries the implication that predictability relationships between words should also follow said structure.

Nevertheless, empirical evidence brought to bear on this matter consists largely of discussion on the merits of specific linguistic examples. Evidence that targets larger parts of the morphological system does exist: for example, Bonami & Strnadová (2019) map out the relationships of form predictability in a subset of related verbs and agent and action deverbal nouns in French, and find that despite the verb supposedly being the base form in this triplet, the action noun is on average just as predictable from the agent noun as it is from the verb, and the action noun is a better predictor of the agent noun than the verb is. Bonami & Guzman Naranjo (2023) find that similar paradigmatic relationships exist for meaning: they train statistical models to predict the distributional vector of a lexeme from the vector of a derivationally related lexeme. They find that the meaning of a lexeme in a given derivational cell is at least somewhat predictable from the meaning of a derivationally related lexeme in a different cell, so implicative relationships of meaning exist in derivational families. Moreover, the meaning of the formal base is not always the best predictor of the meaning of a derivationally related lexeme: they find that lexemes linked by the pattern *Xisme~Xiste* are better predictors of each other’s meaning compared to predicting the meaning of each from the base.

In this talk we compare the rooted tree and the paradigmatic view with the yardstick of cognitive reality. We report on a behavioural experiment to investigate whether speakers are aware of the individual implicative relationships that recent work claims to be structuring derivational paradigms. The experiment aimed to test whether speakers’ mental representation of derivational families resembles more closely the rooted tree view or the paradigmatic view. The paradigmatic conceptualisation of derivational families predicts that speakers would be aware of and exploit all available patterns of predictability, regardless of whether the base is the predictor, or at all involved in the prediction. The rooted tree view would posit that speakers only keep track of relationships of predictability where the base is the predictor.

### 3 Methodology

We performed an acceptability judgement task on French data. We presented speakers with a sentence containing two derivationally related pseudowords and we asked them to rate the acceptability of the second. The more expected the second word is based on the form of the first, the better it was to be rated. Figure 2 shows a sample item.

J’adore le monde de la cationisation. Je veux être	{	<i>catonisateur</i> <i>catonisiteur</i> quand je serai grand. <i>catoniseur</i>
I love the world of ACTION_NOUN. I want to be	{	AGENT_NOUN-1 AGENT_NOUN-2 when I grow up. AGENT_NOUN-3

Figure 2: Sample experimental item, followed by an English translation. Only one of the forms filling the second slot is presented to each participant. The three forms have different levels of predictability conditional on the knowledge that their ACTION NOUN is *catonisation*

Six directed pairs of cells (all permutations of VERB, AGENT NOUN, ACTION NOUN) were chosen for the experiment on the basis of previous work on identifying derivational paradigms

in French (Bonami & Strnadová, 2019). The verb, present in the items only in the infinitive form, was assumed to be the base of both action and agent deverbal nouns by traditional accounts. The cell pairs involve making predictions from the base, towards the base, or between two nonbase cells.

Under a rooted tree view, the predictability of the second word form given the first should only matter when the first form is the verb, since these would be the only cases in which speakers are expected to keep track of predictability relationships. Under a paradigmatic view, word form predictability should have a positive effect in all directed cell pairs.

Participants were shown a video of someone speaking out the items, and signalling which word they should provide an acceptability judgement for. We chose to present stimuli in this way so as not to allow cues from orthography to influence participants' responses. 60 French native speakers were recruited on Prolific.co, and shown 54 crucial items each (9 from each directed cell pair) and 24 distractors (words in inflectional relationships). The same sentence frame could appear with three different pseudoword pairs, of different levels of predictability - the level of predictability of the pseudowords that each sentence appeared with was randomised. Within items for crucial cell pairs, the three levels of predictability were uniformly distributed.

## 4 Results

We fitted a Bayesian mixed effects beta regression to the judgements, predicting them based on the cell pair they instantiate, a phonological well-formedness score of the second form obtained from a separate norming experiment, and the predictability of the second word form based on the first, calculated with the Minimal Generalisation Learner (Albright, 2002; Albright & Hayes, 2003) on data from Démonette (Namer et al., 2019). The conditional effects are pictured below.

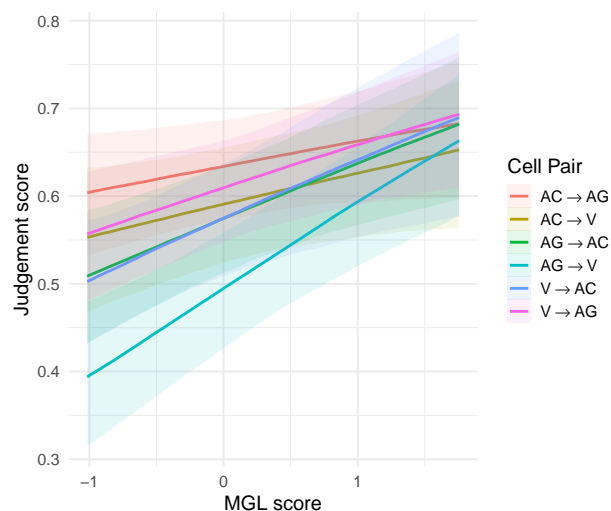


Figure 3: Conditional effect plots of the model

## 5 Discussion

The results fit well with a paradigmatic conception of derivational paradigms within the mental lexicon. Predictability of the second form given the first always has a positive impact on judgement, regardless of the cells involved, in all directions of prediction. A rooted tree view would expect predictability to only have a positive effect when predicting from the base, which

in this case is the verb. Particularly striking is that the case in which predictability matters the most is when predicting from the agent noun towards the verb, which is unexpected under a rooted tree view. It is important to note that the two cell pairs in which the verb is not at all involved also show a positive effect of predictability: a rooted tree view would predict that in this case, if speakers are missing knowledge of the base form and prediction needs to go through the base, speakers would either be lost or would attempt to reconstruct the base, which is problematic since for many of our items where the base was neither the predicted form nor the predictor, the base form was ambiguous.

## References

- Albright, Adam C. 2002. *The identification of bases in morphological paradigms*: University of California, Los Angeles dissertation.
- Albright, Adam C. & Bruce P. Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition* 90. 119–161.
- Aronoff, M. 1976. *Word formation in generative grammar* Linguistic inquiry monographs. Penguin Random House LLC. <https://books.google.fr/books?id=syIXAQAAMAAJ>.
- Bauer, Laurie. 1997. Derivational paradigms. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1996*, 243–256. Dordrecht: Kluwer.
- Becker, Thomas. 1993. Back-formation, cross-formation, and ‘bracketing paradoxes’ in paradigmatic morphology. In *Yearbook of morphology 1993*, 1–25. Springer.
- Bochner, Harry. 1993. *Simplicity in generative morphology*. Berlin, Boston: De Gruyter Mouton. doi:doi:10.1515/9783110889307.
- Bonami, Olivier & Matías Guzman Naranjo. 2023. Distributional evidence for derivational paradigms. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology: theory, methods, evidence*, 219–258. Berlin: De Gruyter.
- Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29. doi:10.1007/s11525-018-9322-6.
- Cotterell, Ryan, Christo Kirov, Mans Hulden & Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics* 7. 327–342. doi:10.1162/tacl\_a\_00271. <https://aclanthology.org/Q19-1021>.
- Hathout, Nabil & Fiammetta Namer. 2014. Discrepancy between form and meaning in word formation: the case of over- and under-marking in french doi:10.1075/cilt.327.12hat.
- Hathout, Nabil & Fiammetta Namer. 2022. Paradis: a family and paradigm model. *Morphology* 32(2). 153–195. doi:10.1007/s11525-021-09390-w.
- Jun, Jongho & Adam C. Albright. 2016. *Speakers’ knowledge of alternations is asymmetrical: Evidence from seoul korean verb paradigms*. Berlin: Cambridge University Press.
- Namer, Fiammetta, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout & Delphine Tribout. 2019. Demonette2 — Une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de TALN*, Toulouse, France. <https://halshs.archives-ouvertes.fr/halshs-02275652/document>.
- Robins, Robert. 1959. In defence of WP. *Transactions of the Philological Society* 58. 116 – 144. doi:10.1111/j.1467-968X.1959.tb00301.x.
- Stump, Gregory. 2019. Some sources of apparent gaps in derivational paradigms. *Morphology* 29. 271–292. doi:10.1007/s11525-019-09339-4.
- Štekauer, Pavol. 2014. Derivational paradigms. In Rochelle Lieber & Pavol Štekauer (eds.), *The oxford handbook of derivational morphology*, 354–369. Oxford: Oxford University Press.



---

# A lexico-paradigmatic analysis of Russian demonyms

*Natalia Bobkova*

*Fabio Montermini*

CLLE, CNRS & Université de Toulouse Jean Jaurès

---

## 1 Introduction

In recent years the study of ethnic terms (henceforth referred to as *demonyms*, also called *ethnonyms* or *gentilics* in various languages, and *katojkonimy* in Russian) has attracted the interest of linguists in several respects, including their lexical status and category, semantic features, variation, as well as their morphological and lexico-semantic properties. As far as derivational morphology is concerned, in particular, demonyms are interesting to study in at least three respects, all of which are related and will be tackled, at some extent, in the present talk:

- they form tight and extremely regular morpholexical networks, which allows to shed light on how these structures interact with derivational morphology (cf. Roché 2008, 2017; Schalchli & Boyé 2018, among others);
- they often make use of a large spectrum of morphological exponents (mainly, but not limited to, affixes), thus constituting an interesting testing ground for approaches to morphological rivalry (cf. Roché & Plénat 2016; Thuilier *et al.* 2021, a.o.);
- they involve peculiar interactions between derivational and inflectional morphology, thus calling into question the frontier between these two domains and to which extent it is (im)permeable (cf. Tuite 1995 on English; Schalchli & Boyé 2017 on French).

In this talk we present a first extensive analysis of a database of ethnic terms (nouns and adjectives) in Russian. The main issues we address concern the relationship between the place (city, region or country) name and its ethnic counterparts, and the network all these words form, in the line of what has been proposed by Schalchli & Boyé (2017) for French. Our analysis is thus carried on in a relational morphology approach, according to which the lexical networks words enter into directly interact with their construction and final form. We also adopt what we call a “constraint-based” model of morphology, according to which word-formation processes correspond to constructions the output form of a derived word as the result of the interaction between a lexeme’s form (a stem, which can possibly undergo various modifications in the derivation process, cf. Roché 2010) and the formal operation linked with a specific construction, which can be viewed itself as a constraint (Author 2 2017).

## 2 The demonymic system of Russian

As in other Slavic languages, and unlike, for instance, the Romance ones, demonymic nouns and adjectives are clearly distinguished lexemes in Russian, each of which follows a specific declensional pattern (for general treatments of demonyms in Russian and Slavic languages in general cf., among others, Akhmetova 2013, 2016; Berezovič 2018). In (1) we present some sets of lexemes including a toponym (country / region / city name), an ethnic adjective (roughly meaning ‘related to the place X’), a masculine and feminine ethnic noun (referring to the inhabitants of a country / region / city).

(1) <sup>1</sup> a.	Burjatij(a) ('Buryatia')	burjatsk(ij) <sub>A</sub>	burjat <sub>N,M</sub>	burjatk(a) <sub>N,F</sub>
b.	Volgograd	volgogradsk(ij) <sub>A</sub>	volgogradec <sub>N,M</sub>	volgogradk(a) <sub>N,F</sub>
c.	Kirov	kirovsk(ij) <sub>A</sub>	kirovčanin <sub>N,M</sub>	kirovčank(a) <sub>N,F</sub>
d.	Ostrov	ostrovsk(ij) <sub>A</sub>	ostrovič <sub>N,M</sub>	ostrovičk(a) <sub>N,F</sub>
e.	Zelenogorsk	zelenogorsk(ij) <sub>A</sub>	zelenogorec <sub>N,M</sub>	zelenogork(a) <sub>N,F</sub>

Examples (1a-d) illustrate the most frequent types of formal relations observed in Russian demonyms. More in detail, as these examples show, the ethnic adjective is derived by means of the suffix *-sk-* without exceptions.<sup>2</sup> As far as nouns are concerned, the situation is more complex. As other languages, Russian possesses a bunch of simple ethnic (masculine) nouns from which toponyms are created (1a). More often, however, masculine inhabitant nouns are constructed by one of the three suffixes *-c*, *-anin* or *-ič* and their variants (cf. below). Table 1 presents the proportion of masculine demonymic nouns in our database, distinguishing between Russian and foreign toponyms (cf. below for details on the database).

	<i>-c</i> (type 1a)	<i>-anin</i> (type 1b)	<i>-ič</i> (type 1c)
Russian	663 (69,94%)	277 (29,22%)	8 (0,84%)
Foreign	297 (95,50%)	14 (1,53%)	–

**Table 1: distribution of the main construction types of masculine denonymic nouns in the database**

Feminine inhabitant nouns, on their turn, are consistently constructed by means of the suffix *-k* (or a variant of it),<sup>3</sup> which attaches either to the 'bare' form of the toponym (cf. 1b, e), or to the masculine noun (cf. 1 a, c, d). Moreover, the situation is made more complex by the fact that many Russian city names include themselves suffixes, such as *-sk* (1e), which is the outcome of a further derivation from the corresponding adjective (cf. Cexanovič 2007).

### 3 Database construction

For our research we collected a database of Russian demonyms from various sources, namely the list of Russian cities provided in the Russian Wikipedia<sup>4</sup> and Babkin (ed.) (1964) for demonyms of Russian place names and the Russian Wikipedia, as well as other Internet sources for demonym of foreign place names. Since Internet sources, and in particular Wikipedia, are not always compiled according to strict lexicographic criteria, to be included each demonym had to appear at least once in the Russian National Corpus,<sup>5</sup> or, in alterna-

<sup>1</sup> Glosses are provided only for country or region names; when a gloss is lacking, the geographic name designates a Russian city. Brackets in the representation of words indicate the inflectional suffix of the citation form ((masculine) nominative singular).

<sup>2</sup> *-sk-* is one of the three main denominal adjectival suffixes in Russian, along with *-n-* and *-ev-/-ov-* (cf. Zemskaja 2015; Kustova 2018; Autor 1 & Autor 2 in press).

<sup>3</sup> Both *-c* and *-k* possess an allomorph displaying an extra vowel when they appear in suffixless inflected forms (cf. *vinogradec<sub>M,NOM,SG</sub>*, *vinogradok<sub>F,GEN,PL</sub>*) (on vowel/∅ alternations in Russian cf. Sims 2017 among others).

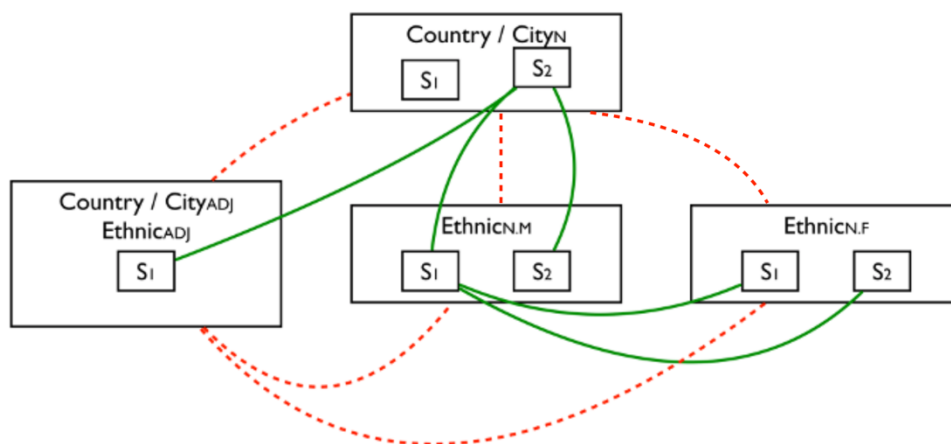
<sup>4</sup> [https://ru.wikipedia.org/wiki/Список\\_городов\\_России](https://ru.wikipedia.org/wiki/Список_городов_России).

<sup>5</sup> <https://ruscorpora.ru/>.

tive, had to have at least one attestation in discourse in a Google search. This allowed us to gather a database of 3,523 demonyms from 1,133 Russian city names and 915 demonyms from 279 foreign place (mainly country) names.<sup>6</sup> The interest of distinguishing two different databases was, among others, the fact that it allows us observing the different treatment these demonyms receive, and to have hints on the actual productivity of each type. For instance, if we consider that demonyms from foreign toponyms have globally entered the lexicon more recently than those from Russian names, we can conclude that the suffix *-c* is overwhelmingly the most productive for the derivation of masculine names.

## 4 Analysis

On the basis of the collected database we propose a lexico-paradigmatic analysis of demonym formation in Russian. In our analysis, the lexico-semantic properties are partly disconnected from the formal ones. In particular, the former connect lexemes (or rather “lexemes”, as in Schalchli & Boyé’s 2018 analysis), whereas the latter connect stems stored in the lexical representation of lexemes. The overall schema of Russian demonym construction is presented in Figure 1, where red dotted lines represent lexico-semantic links, and green lines represent formal links.



**Figure 1: global scheme of Russian demonym construction**

In particular, we consider that place names in Russian have a two-stem stem space, including a ‘hidden’, exclusively derivational, stem. This accounts for systematic allomorphies encountered both with adjectives and nouns. These include derivational variants of Slavic origin (like in *Dn(o) / dnovsk(ij) / dnovec*) and adaptations of ethnic affixes of foreign origin (like in *Korsika* ‘Corsica’ / *korsikansk(ij) / korsikanec*). Moreover, feminine nouns are also constructed formally (but not semantically) on the default stem of masculine nouns. Finally, we consider that, like in other languages (cf. Roché 2008; Schalchli & Boyé 2018 on French), the ethnic adjective, although it is formally linked only to the place name, is semantically linked to both the latter (meaning ‘related to the country / city X’) and to the ethnic names (meaning ‘related to the inhabitants of X’). In particular, we provide numerical evidence for

<sup>6</sup> The number of actually considered forms is higher than what expected for the number of toponyms considered because some places names display variation and more than one form are attested.

the different cases encountered, and for the fact that the schema proposed allows accounting for the great majority of them and for the variation observed.

## References

- Akhmetova, Marija V. 2013. Variantnost' nazvanij žitelej (po materialam èlektronnoj bazy SMI "Integrum"). *Vestnik RGGU* 8. 86-105.
- Akhmetova, Marija V. 2016. Strategii nominacii i "jazykovaja politika": nazvanija žitelej gorodov. *Labirint* 5. 76-85.
- Babkin, Aleksandr M. (ed.) (1964). *Slovar' nazvanij žitelej (RSFSR)*. Moskva: Sovetskaja Ènciklopedija.
- Berezovič Elena L. 2018. Slavjanske ottoponimičeskie nazvanija žitelej: asociativno-derivacionnaja i frazeologičeskaja semantika. In: Svetlana M. Tolstaja (ed.), *Slavjanskoe jazykoznanie. XVI Meždunarodnyj s'ezd slavistov. Belgrad 20-27 avgusta 2018 g.* 7-34. Moskva: Institut Slavjanovedenija Ran.
- Cexanovič, Marianna A. 2007. Suffiks -sk- v rusškoj toponimii i problema naloženija morfem v ottoponimičeskix prilagate'lnyx. *Vestnik PSTGU. III Filologija* 3.9. 14-22.
- Kustova, Galina I. 2018. Prilagatel'nye. In Vladimir A. Plungjan & Natal'ja M. Stojnova (eds.), *Materialy k korpusnoj grammatike russkogo jazyka, vyp. III: Časti reči i leksiko-grammatičeskie klassy*, 40-107. Sankt-Peterburg: Nestor-Istorija.
- Roché, Michel. 2008. Structuration du lexique et principe d'économie : le cas des ethniques. In Jacques Durand, Benoît Habert & Bernard Laks (eds.), *Actes du Congrès mondial de linguistique française*, 1571-1585. Paris: Institut de Linguistique Française.
- Roché, Michel. 2010. Base, thème, radical. *Recherches linguistiques de Vincennes* 39. 95-134.
- Roché, Michel & Marc Plénat. 2016. De l'harmonie dans la construction des mots français. In Franck Neveu, Gabriel Bergounioux, Marie-Hélène Côté, Jean-Marie Fournier, Linda Hriba, Sophie Prévost (eds.), *5<sup>e</sup> Congrès mondial de linguistique française*. Paris: Institut de Linguistique Française.
- Schalchli, Gauvin & Gilles Boyé. 2018. Paradigms and syncretism in derivation: The case of ethnics in French. *Lingue e linguaggio* XVII.2. 197-215.
- Sims, Andrea. 2017. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics* 25.2. 143-182.
- Thuilier, Juliette, Delphine Tribout & Marine Wauquier. 2021. Affixal rivalry in demonym formation. Paper presented at the on-line workshop on affix rivalry, 19 March 2021.
- Tuite, Kevin. 1995. The declension of ethnonyms in English. *Proceedings of the Twenty-First Annual Meeting of the Berkeley Linguistics Society* [https://journals.linguisticsociety.org/proceedings/index.php/BLS/article/view/1420/0].
- Zemskaja, Elena A. 2015. *Jazyk kak dejatel'nost': Morfema, slovo, reč'*. Moskva: Flinta.

# **POSTER PRESENTATIONS**



---

# Sémantique des adjectifs dénominaux suffixés en *-u* et construits à partir d'un « nom d'élément du corps »

Thomas Bertin

Université de Brest (UBO)  
LaTIM & Département d'orthophonie

Nicole Bessière

Université de Rouen  
Département des Sciences du langage

---

## 1 Introduction

L'étude porte sur une petite classe d'adjectifs du français construits à partir d'un nom désignant un *élément* du corps humain ou animal ( $N_{ec}$ ) suffixé en *-u* (p. ex. *ventru*, *poilu* ou *pattu*). L'objectif principal est d'en proposer une analyse sémantique articulant certains acquis concernant, d'une part, le suffixe *-u* (Mélis-Puchulu 1991 ; Aurnague & Plénat 1997, 2007 ; Temple 2002) et, d'autre part, les noms de *parties* du corps humain ( $N_{pc}$ ) (Bertin 2018, 2021) qui forment une sous-classe de celle des  $N_{ec}$ . Secondairement, nous confrontons cette analyse aux usages.

## 2 Délimitation du corpus d'étude

Le corpus d'étude est constitué à partir de la base de données morphologiques du projet *Demonext*. Celle-ci recense **42 adjectifs** dénominaux suffixés en *-u*. Seuls sont retenus ceux construits à partir d'un *nom d'élément du corps*.

### 2.1 Nom d'élément du corps ( $N_{ec}$ )

Les noms de *partie du corps humain* ( $N_{pc}$ ) sont des méronymes emblématiques. Ils instancient en effet la relation *partie-tout* prototypique dite *composant-assemblage* (Cruse 1986 ; Winston *et al.* 1987 ; Aurnague 1996). Une définition linguistique des  $N_{pc}$  a été proposée (Bertin 2018) qui prend en compte (i) un critère de **dénomination** (Kleiber 1984) – par exemple, *tête* dénomme une partie du corps humain au contraire de *citron*, malgré des énoncés du type *il a rien dans le citron* – et (ii) un critère de **possession inaliénable** (Mostrov 2015) – on peut ainsi opposer *Paul a mal à la/?sa tête* à *Paul a mal à ?la/sa verrue*. Sur cette base, le nom *tête* est considéré comme un  $N_{pc}$  au contraire de *citron* ou *verrue*.

La classe des *noms d'élément du corps* ( $N_{ec}$ ) comprend les  $N_{pc}$  mais elle est plus large. D'une part, y sont intégrés des noms dénommant des parties du corps **animal**. En fait, la frontière s'avère poreuse du point de vue linguistique. Ainsi, des noms sont indifféremment utilisés pour l'être humain ou l'animal (cf. *le dos/la tête de l'homme/l'âne*). Par ailleurs, des noms dénommant des parties du corps animal sont utilisés pour désigner des parties du corps humain dans des énoncés comme *Paul s'est cassé la gueule*, *Joël s'en met plein la panse*, *Léa en a plein les pattes*.

D'autre part, le critère de possession inaliénable est écarté. Des noms comme *verrue* ou *poil* – désignant un référent *produit* par le corps plutôt que *partie* du corps – sont considérés comme des  $N_{ec}$  (bien qu'ils ne soient pas des  $N_{pc}$  au sens précédent).

### 2.2 Corpus d'étude

Parmi les 42 adjectifs évoqués précédemment, **24 sont construits à partir d'un  $N_{ec}$**  (soit 57%). Le poster les présentera tous. Certains sont construits de façon transparente (p. ex. *fess-u*), d'autres à partir d'un allomorphe (p. ex. *cheveu/chevel-u*), d'autres encore à partir d'un nom issu d'un fonds lexical ancien (p. ex. *lippe/lipp-u*).

### 3 Sémantique des $N_{ec} + u$

Selon Mélis-Puchulu (1991), tout adjectif dénominal a un sens *relationnel* : il établit une relation entre le nom de base à partir duquel il est construit ( $N_b$ ) et le nom recteur dont il dépend syntaxiquement ( $N_r$ ).

Pour le suffixe *-u*, Aurnague et Plénat (1997 : 18) parlent d'une « relation de possession » entre les référents du  $N_r$  et du  $N_b$ . Ainsi, dans *un homme ventru*, l'adjectif *ventru* établit une relation de possession entre les référents de *homme* ( $N_r$ ) et de *ventre* ( $N_b$ ). Corollairement, le *TLFi* précise que *-u* signifie « qui possède, qui est caractérisé par le radical » et Dubois et Dubois-Charlier (1999 : 148) le glosent par « qui a / est muni de / possède » le référent du  $N_b$ . Ces observations méritent d'être précisées pour la suffixation avec un  $N_{ec}$ .

Si les  $N_{ec}$  n'instancient pas tous une relation *partie-tout* prototypique (cf. *barbiche/chair*), ils relèvent bien de ce que Aurnague et Plénat (1997, 2007) appellent une « relation d'attachement habituel ». Par ailleurs, ils ne dérogent pas aux contraintes postulées par ces deux auteurs et Temple (2002 : 209-210) : (i) *exclusion des artefacts* et (ii) *exigence de saillance visuelle*. De fait, les  $N_b$  du corpus réfèrent (i) aux *parties biologiques* d'êtres animés (cf. 4.1 pour les emplois métaphoriques) mais (ii) pas à des organes *internes*. Les adjectifs *ossu* et *pansu* ne sont pas des contre-exemples car ils soulignent précisément à quel point les os et la panse sont saillants (i.e. visibles de l'extérieur).

Nous défendons que, combiné à un  $N_{ec}$ , le suffixe *-u* met en jeu la **notion d'excès**. Dans cette perspective, nous partons d'une description des emplois QUALITÉ des  $N_{pc}$  (Bertin 2018, 2021). En effet, dans certains contextes, un  $N_{pc}$  renvoie à une qualité psychologique (p. ex. *Paul a du cœur* 'de la générosité') ou physiologique (p. ex. *Paul a du ventre* 'un gros ventre'). Contingente (au contraire de la partie du corps, elle n'est pas nécessaire) et relativement permanente (au contraire de l'état, elle n'est pas passagère), la qualité est susceptible de caractériser un référent, notamment animé (cf. Flaux & Van de Velde, 2000 : 88). Or, avec Leeman (2004 : 156), on note que le partitif peut évoquer à lui seul une « grande quantité ». Ainsi, *avoir du cœur* c'est 'avoir (beaucoup) de générosité' et *avoir du ventre* c'est 'avoir trop de ventre'.

Quoique la construction partitive apparaisse assez aléatoire avec les  $N_b/N_{ec}$  (*avoir de la barbe/\*bosse*), on retrouve la notion d'excès dans des exemples comme *un homme barbu/bossu* 'qui a une barbe/bosse' (attribut non nécessaire présent en *surplus*) ou *un homme poilu/ventru* 'qui a beaucoup de poils/ventre' (attribut nécessaire présent en *abondance*). Cette analyse fait écho aux descriptions lexicographiques du *TLFi* qui évoque la « valeur intensive » du suffixe *-u* et du *Dictionnaire étymologique et historique* (Larousse) qui le classe comme « augmentatif ».

Elle trouve confirmation dans la comparaison avec les suffixes *-eux* et *-é* desquels *-u* est parfois rapproché (Mélis-Puchulu 1991 ; Dubois & Dubois-Charlier 1999). Ainsi, on peut opposer *le système pileux/??poilu* à *Jeanne est ??pileuse/poilue* ou *un régime carné/??charnu* à *une bouche ??carnée/charnue*. Les adjectifs *pileux* et *carné* fonctionnent comme de simples catégorisants sans valeur qualifiante – strictement « relationnels » au sens de Bartning & Noailly (19993) et Noailly (1999 : 22). Or, si les adjectifs *poilu* et *charnu* ne sont certes pas dépourvus de toute fonction catégorisante ('a des poils' et 'est faite de chair'), ils véhiculent une valeur qui relève de l'excès : 'a des poils en excès' et 'a beaucoup de chair'. Corollairement, on oppose *osseux* et *ossu* (*tumeur osseuse* 'de l'os' vs. *chien ossu* 'avec de gros os') ou *denté* et *dentu* (*roue dentée* 'avec des dents' vs. *visage dentu* 'qui se singularise par des grandes/grosses dents') même si *ossu* et *dentu* sont moins courants.



## 4 Usage des $N_{ec} + u$

Des sondages dans *frWaC* permettent de distinguer des adjectifs très courants (des milliers d'occurrences de *barbu*, *charnu* ou *têtu*), courants (des centaines d'occurrences de *bourru*, *joufflu* ou *velu*), rares (41 occurrences de *barbichu*, 36 de *fessu* et 15 de *lippu* – recensées après filtrage des doublons, noms propres, termes de zoologie...) ou très rares (9 occurrences de *dentu*, 5 de *mamelu*, 1 de *dentu* et 0 de *membre*). Sans négliger cette dimension quantitative, l'usage sera préférablement abordé au prisme du fonctionnement sémantique.

### 4.1 Emplois métaphoriques

Avec les  $N_{pc}$ , la possibilité est ouverte à des emplois *psychologiques* (p. ex. *Paul a du nez* 'de l'intuition') plutôt que proprement *physiologiques*. On peut considérer ces emplois comme métaphoriques au sens où le  $N_{pc}$  (ici, *nez*) ne renvoie plus à la partie du corps mais, par analogie, à une qualité psychologique. On retrouve ce fonctionnement avec l'expression familière *avoir des couilles* dont le sens 'avoir du courage/de l'audace' correspond à celui de *couillu* apparu récemment (Aurnague & Plénat, 2007 : 64) mais devenu courant. Plus courant encore, *goulu* 'vorace' a cependant un fonctionnement moins transparent (\**avoir de la goule* ; *avoir de la gueule* s'emploie dans un sens différent : 'être classe'). Quant à *têtu*, s'il a pu, dans un état précédent de langue, signifier 'avoir une grosse tête', il signifie désormais 'obstiné' (sans lien avec un potentiel ??*avoir de la tête* relevé chez Annie Ernaux).

Enfin, les adjectifs *bourru* et *cornu* se prêtent à deux interprétations : physiologique (*bras/torse bourru* (vieilli) ; *une bête cornue*) ou plus abstraite (*Paul est bourru* 'farouche' ; *Paul est cornu/cocu/a des cornes* 'victime d'une infidélité amoureuse'). Notons que l'adjectif \**cœur* ('avoir du courage/de la noblesse') est sorti de l'usage au profit de *avoir du cœur* (qui évoque désormais plus souvent la générosité).

Indépendamment du  $N_b$ , le fonctionnement métaphorique peut s'imposer par la présence d'un  $N_r$  au référent non animé. Dans *frWaC*, on trouve *une grappe de petite baies charnues* ou *ce récipient pansu*. Ici, le sens métaphorique rend la notion d'excès d'autant plus explicite.

### 4.2 Ressort discursif du suffixe -u

La notion d'excès peut être simplement mobilisée pour formuler un énoncé s'apparentant à une définition : *Le wavy-Coated Retriever [...] était un chien ossu* ['aux gros os solides'] *et très vigoureux, avec une tête large* (*frWaC*). Cependant, dans nombre d'énoncés, l'excès confine à l'excentricité voire à la subversion (notamment *via* des procédés d'accumulation) : *elle est belle, sérieuse, adroite, jambeuse, fessue, pétante de poitrine* ou *mais tout à coup, de cette nuit, surgissent trois êtres ventrus, fessus, bossus* (*frWaC*). Parfois, l'excès est mis au service de l'évocation d'une forme de monstruosité : *une peuplade de personnages caricaturaux, ivrognes, goinfres et libidineux, anguleux et mamelus, grotesques, mesquins, cacochymes* (*frWaC*).

Les néologismes mobilisent également la notion d'excès. En voici, deux exemples : *Pedro, ancien talonneur rugueux et cuissu* ['aux fortes cuisses'] *du Sporting Mirandolais dans les années 1960* (La dépêche du Midi – Sept. 2017) ; *Il préférait à une bergère de chanson, c'était visible, quelque grasse fille hanchue* ['aux larges hanches'] (G. Guevremont, *Le survenant* – 1945). Ces deux adjectifs sont également relevés par Aurnague et Plénat (2007 : 65-66).

De telles données amènent à anticiper de possibles néologismes. Ainsi, *veinu* paraît plus probable qu'*artéru* (le nom *veine* désigne un élément du corps perceptible visuellement et susceptible de foisonner, sur une main par exemple).

## 5 Conclusion

Plus que strictement relationnels, les adjectifs du type  $N_{ec} + u$  véhiculent une valeur d'excès qui fait écho aux emplois QUALITÉ des  $N_{pc}$ . Celle-ci est attestée dans les emplois concrets (*cet homme est poilu* 'a beaucoup de poils') mais aussi dans les emplois métaphoriques (*cet enfant est goulu* 'est vorace' ou, attesté dans *frWaC*, *des kouglofs joufflus*) comme au détour d'énoncés singuliers, parfois truculents, intégrant des néologismes à l'occasion.

## Références

- Aurnague, Michel. 1996. Les noms de localisation interne – Tentative de caractérisation sémantique à partir de données du basque et du français. *Cahiers de Lexicologie* 69, 159–192.
- Aurnague, Michel & Marc Plénat. 1997. Manifestations morphologiques de la relation d'attachement habituel. *Sillexicales* 1, 15–24.
- Aurnague, Michel & Marc Plénat. 2007. Contraintes sémantiques et dérivation en *é* : attachement habituel, naturalité et dissociation intentionnelle. *Carnets de Grammaire* 16 – *Rapports internes de CLLE-ERSS* (Toulouse).
- Bartning, Inge & Michèle Noailly. 1993. Du relationnel au qualificatif : flux et reflux. *L'information grammaticale* 58, 27–32.
- Bertin, Thomas. 2018. *La polysémie des noms de parties du corps humain – Analyse sémantique de artère, bouche, cœur, épaule et pied*. Thèse de l'université de Rouen.
- Bertin, Thomas. 2021. Les noms de parties du corps humain en français : proposition de classement des acceptions. *Cahiers de lexicologie* 119, 73–100.
- Cruse, Alan D. 1986. *Lexical semantics*. Cambridge University Press.
- Demonext (projet). Base de données morphologiques en ligne. <https://www.demonext.xyz/> [consulté le 2 mars 2023].
- Dubois, Jean & Françoise Dubois-Charlier. 1999. *La dérivation suffixale en français*. Paris : Nathan Université.
- Dubois, Jean, Henri Mitterrand & Albert Duazat. 2011. *Dictionnaire étymologique et historique du français*. Paris : Larousse.
- Flaux, Nelly & Danièle Van de Velde. 2000. *Les noms en français : esquisse de classement*. Paris/Gap : Ophrys.
- Kleiber, Georges. 1984. Dénomination et relations dénominatives. *Langages* 76, 77–94.
- Leeman, Danielle. 2004. *Les déterminants du nom en français*. Paris : PUF.
- Mélis-Puchulu, Agnès. 1991. Les adjectifs dénominaux : des adjectifs de "relation". *Lexique* 10, 33–60.
- Mostrov, Vassil. 2015. L'être humain et la relation partie-tout. In Wiltrud Mihatsch & Catherine Schnedecker (eds), *Les noms d'humains : une catégorie à part ?*, 115–146. Stuttgart : Verlag.
- Noailly, Michèle. 1999. *L'adjectif en français*. Paris/Gap : Ophrys.
- Temple, Martine. 2002. Métaphore et mots construits : éclairages réciproques. *Verbum* XXIV-3, 207–227.
- Trésoir de la langue française informatisé (Le) – ATILF (Nancy)*. 2004. Dictionnaire en ligne. <http://atilf.atilf.fr/tlf.htm> [consulté le 12 mars 2023].
- Winston, Morton E., Roger Chaffin & Douglas Hermann. 1987. A taxonomy of part-whole relations. *Cognitive Science* 11, 417–444.

# The binary vs. privative status of verbal inflectional morphology: The case of Germanic

Concha Castillo  
University of Málaga

## 1 Background and questions to pose from a DM perspective

I argue that the binary opposition [+/-past] entails that  $T_{\text{past}}$  contrasts with  $T_{\text{pres}}$  in computing one more  $\tau$  (or tense)–feature in the morpho-syntax and exhibiting one more Vocabulary Item (or marker) in the morpho-phonology. This used to be the case for all Germanic languages in their old periods but is no longer the case for Present Day English or Mainland Scandinavian.

From a broad formalist point of view, the (non-periphrastic) Present tense and Past tense in Germanic languages appear to fit particularly well with the *binary* specification [+/-past], since a concrete marker or segment –namely, the dental segment– expones exclusively in Past forms and can thus intuitively be used as a criterion to characterize these as *marked* forms as compared to the Present. Identifying the Past as the morpho-syntactically *marked* form requires nevertheless to account in an exhaustive way also for the Present.

From the perspective of *Distributed Morphology* (DM), for which morphological markers or, the same, *Vocabulary Items* (VI's), are the (morpho-phonological) output of the computation of (morpho-syntactic) features, Present and Past forms in a language like Present Day English (PDE) are equally characterized as [+/-past]. One aspect that DM highlights (Halle & Marantz 1993) is the mismatch between morpho-syntax and morpho-phonology that could be argued to exist between Present and Past forms in the language since, aside from the stem, the VI that is overtly realized is the output of a tense feature ( $\tau$ ) in the case of the Past (the cited dental segment) while it is the output of an agreement feature ( $\varphi$ ) in the case of the Present (the segment typically or traditionally referred to as *subject agreement ending*): note *deem-s* vs. *deem-ed*. The way that this mismatch is accounted for is by invoking a process of *fusion*, which would be additionally preceded by *impoverishment* in the case of the Present. See Table 1 below. Incidentally, in order to save space in this abstract, reference is only to regular pasts for all languages cited; further, allomorphy of the dental segment is not relevant for the argumentation and is therefore obviated. I do not use here phonetic transcriptions.

**Table 1. Segmentation for Present and Past forms in PDE previous to *fusion* (DM generalized account)**

	Present Indicative of <i>deem</i>	Past Indicative of <i>deem</i>
1sg	deem- $\emptyset$ - $\emptyset$ <b>STEM-<math>\tau</math>feature-<math>\varphi</math>feature</b>	deem-ed- $\emptyset$ <b>STEM-<math>\tau</math>feature-<math>\varphi</math>feature</b>
2sg	deem- $\emptyset$ - $\emptyset$	deem-ed- $\emptyset$
3sg	deem- $\emptyset$ -s	deem-ed-s
Pl	deem- $\emptyset$ - $\emptyset$	deem-ed- $\emptyset$

In effect, in order for the morpho-syntax to be (initially) symmetric, Halle & Marantz (1993) postulate a mechanism of impoverishment as applying in the Present: note the  $\emptyset$ –segment in medial position, which means that there is a  $\tau$ –feature for the Present, in a symmetric way to the Past, though this morpho-syntactic feature is bound to have no morpho-phonological realization. Turning to the Past, the segment that would correspond to the  $\varphi$ –feature is added, this time in a symmetric way to the Present. Subsequently, the very impossibility of *\*he/she deem-ed-s* leads to positing *fusion*, which consists in that only one VI will be inserted for both  $\tau$ –features and  $\varphi$ –features, the other being obliterated: the VI

or output of corresponding  $\tau$ -features is cancelled out for the Present and the VI or output of corresponding  $\varphi$ -features is cancelled out for Past forms. Incidentally, it must be clarified that it is fusion of *heads* that the authors specifically refer to: as is widely known, Early Minimalism inherits the hierarchical sentence structure of the GB period where Agr(eement)P(hrase) and T(ense)P(hrase) are both canonical projections, and where the checking or computation of agreement or  $\varphi$ -features ([person] and [number]) corresponds to the Agr head and that of tense or  $\tau$ -features (above-cited [+/-past]) corresponds to the T head. The subsequent generalized consensus in the literature on the rejection of an Agr projection proper in the process of derivation of syntactic structures leads to the likewise generalized account of T as the head in charge of the computation of  $\tau$ -features and  $\varphi$ -features (Chomsky 2000, 2001; Pesetsky & Torrego 2004/2007 or quite recently e.g. Bjorkman & Zeijlstra 2019). Having said this, the core of the analysis in Halle & Marantz (1993) remains: that is, only one type of feature –either  $\tau$ -features or  $\varphi$ -features– expones in the English morpho-phonology, at the cost of the other.

**Table 2. Segmentation for Present and Past forms in PDE after fusion**

	Present	Past
1sg	deem- $\emptyset$	deem-ed
2sg	deem- $\emptyset$	deem-ed
3sg	deem-s	deem-ed
Pl	deem- $\emptyset$	deem-ed

I would like to argue that it is necessary to raise the following questions or issues in connection with the account in Table 1:

(1) It is not clear in what sense it is to be concluded that a morpho-syntactic [+/-past] feature is available in PDE. That is, in what sense are Past forms *marked*, rather than Present forms? Now, maybe it is implicitly assumed that it is *exclusively contrastive* values and not *marked* values that are involved, which should mean that Present and Past forms are the result of the computation of two *privative* (or also *unary*) features, rather than a *binary* feature, a description that is actually the one I defend for PDE in this proposal. But the focus must be put on the account or analysis proper in Table 1. And in this sense, I would like to argue the following.

(2) It does not seem to be explanatory to start by assuming a *symmetric* status for the Present and the Past when a situation of *asymmetry* can be at stake, that is a situation where the number of morpho-syntactically active features can be bigger for one of the elements in the relevant opposition, and as a result of this, the number of realized segments or VI's.

(2bis) Regarding specifically impoverishment (or otherwise a rule of *obliteration*, as in Arregui & Nevins (2012) or the *pruning* of a T head, as in Embick (2015)), this can be indeed an impeccable mechanism for other situations, but for it to be presented as the cause of a non-realizational *default* appears to be fully *ad hoc*.

(3) It does not seem to be explanatory to assume an -s marker for 3 person sg in the Past (Table 1), it being the case that there are (Germanic) languages where subject agreement markers are different for the Present and the Past. Incidentally, as I defend in my research, this is the case for languages descending from Proto-Indo-European in general and it plays a major role in the account I defend here (see (B) in Section 2 below).

(4) If the account in Table 1 is applied to a language like German (or also Icelandic, or Frisian), and it being the case that the segmentation for Past forms is as in *kauf-te- $\emptyset$*  ('I bought') (as generally assumed), then it would be so expected that impoverishment is implemented on Present forms, with a result as in *kauf- $\emptyset$ -e* ('I buy'). I do not think this is explanatory because of the reasons in (2) and (2bis) above, and because of (B) in Section 2 below.

(5) If the account in Table 1 is applied to a language like Danish (or also Swedish, or Norwegian), there is the additional issue of resolving first whether the segmentation for Past forms is as in *hØr-t-e* or otherwise as in *hØr-te* ('I...heard'). Then, on the cited symmetric account (which, as I say, does not seem to be explanatory enough) Present forms will either be *hØr-Ø-er* (with impoverishment) or *hØr-er* ('I...hear') (without it).

## 2 Present proposal

I would like to argue that for [+/-past] to be the expression of *morpho-syntactic binarity* in Germanic languages (and, in ongoing research, in languages descending from PIE) entails that Past forms are *marked* in the sense that *one more formal feature* is active in their computation as compared to Present forms, and *one more segment or VI* is spelled out in the morpho-phonology. Within Germanic, I argue that languages like German, Icelandic or Frisian do compute the cited *binary*  $T_{\text{pres}}/T_{\text{past}}$ , whereas PDE on the one hand and Present Day Mainland Scandinavian on the other do compute a *privative*  $T_{\text{pres}}$  and a *privative*  $T_{\text{past}}$ . The account defended is both cross-linguistic and diachronic. Parting from (1)–(5) above, the basic line of argumentation is as follows:

(A) The account defended is both cross-linguistic and diachronic, since it is the case that Present Day German (or Icelandic...) exhibits a segmentation of VI's that can be considered to be identical to the segmentation of all Germanic languages in their old periods.

(B) The cited segmentation consists, as regards Past forms, of the (widely-known) dental marker or VI which can be arguably uncontroversially be analyzed as the output of a  $\tau$ -feature with the interpretation [past], plus the so-called subject agreement ending which, in a crucial way, *I defend must be analyzed also as a  $\tau$ -feature*, though it is a  $\tau$ -feature that is a kind of *portmanteau* since it combines  $\varphi$ - and  $\tau$ -interpretation. I refer to this feature as an *AgrT-feature*. The content of  $\varphi$ -interpretation is [person] and [number] as standard. The content of  $\tau$ -interpretation, which is why it must significantly be analyzed as a proper  $\tau$ -feature, is [*morphological distinctiveness both within and across the Present and the Past*] (see (3) in Section 1 above). This takes us to Present forms, which consist (aside from the stem) of just this subject agreement ending, that is an *AgrT-feature* with the content [present]. Past forms result therefore from the computation of a double(d)  $\tau$ -licensing as compared to Present forms. Consider the unanimous segmentation to the left of the arrow for all cases in Table 3 below. (I assume general tenets of DM relative to the *Subset Principle*, the *Elsewhere condition* and also *Fusion* – though no *Impoverishment*. And I assume core principles of the *Agree* framework (Chomsky 2000, 2001; Pesetsky & Torrego 2004/2007) in connection with the licensing of  $\tau$ -features and  $\varphi$ -features.)

**Table 3. Diachronic development of morpho-syntactic features on the present account**

<b>English</b>	→ diachronic change: first half of 18 <sup>th</sup> cent.
(Present) stem - [+present]AgrT-feature	→ stem - [present] $\tau$ -feature
(Past) stem - [-present]AgrT-feature - [past] $\tau$ -feature	→ stem - [past] $\tau$ -feature
<b>Danish, Swedish, Norwegian</b>	→ diachronic change: first half of 17 <sup>th</sup> cent.
(Present) stem - [+present]AgrT-feature	→ stem - [present] $\tau$ -feature
(Past) stem - [-present]AgrT-feature - [past] $\tau$ -feature	→ stem - [past] $\tau$ -feature
<b>German, Icelandic, Frisian</b>	
(Present) stem - [+present]AgrT-feature	→ no diachronic change
(Past) stem - [-present]AgrT-feature - [past] $\tau$ -feature	→ no diachronic change

(C) The evidence that I provide for the active computation of the *AgrT-feature* in the old stages of English and Mainland Scandinavian and its demise in the first half of the 18<sup>th</sup> cent. and the 17<sup>th</sup> cent., respectively, relates to the phenomenon of so-called V-to-T movement: it is when  $T_{\text{pres}}$  and  $T_{\text{past}}$  stop contributing a binary opposition (in the morpho-syntactic way defended here) that these languages stop being V-to-T and become V-*in situ*. Note the identical segmentation to the right of the arrow for these

languages, irrespective of the major role played by the  $\emptyset$ -VI in English as opposed to Danish. The last two-column division in each of the Tables below is one where the so-called subject agreement endings no longer interpret [*morphological distinctiveness both within and across the Present and the Past*].

**Table 4. Historical development of the morpho-phonology of the [AgrT]-feature for English**

OE		Late ME (c. 1400)		EMnE(c.1500→1700)		1700→PDE	
Present	Past	Present	Past	Present	Past	Present	Past
1sg -e	-e	- $\emptyset$ /-e	-e/- $\emptyset$	- $\emptyset$	- $\emptyset$	- $\emptyset$	-_____
2sg -e(st)	-(e)st	-st	-st	-st	-st	- $\emptyset$	-_____
3sg -eþ	-e	-th/-s	-e/- $\emptyset$	-th/-s	- $\emptyset$	-s	-_____
Pl -aþ	-on	-n/-s/-th	-e(n)	- $\emptyset$	- $\emptyset$	- $\emptyset$	-_____

**Table 5. Historical development of the morpho-phonology of the [AgrT]-feature for Danish**

Middle Danish (1300)		Early Modern Danish (1500)		1600 → PD Danish	
Present	Past	Present	Past	Present	Past
1sg -e(r)	-e	-er	-e	-er	-e
2sg -er	-e/-(s)t	-er	-e	-er	-e
3sg -er	-e	-er	-e	-er	-e
1pl -e/-um	-e//e/-um	-e	-e	-er	-e
2pl -e	-e	-e	-e	-er	-e
3pl -e	-e	-e	-e	-er	-e

## 2.1 A more detailed description of the historical case for English (...)

## 2.2 A more detailed description of the historical case for Mainland Scandinavian (...)

## 2.3 A more detailed description of the historical case for German, Icelandic, Frisian (...)

## References

- Arregui, Karlos & Andrew Nevins. 2012. *Morphotactics. Basque auxiliaries and the structure of Spellout*. Dordrecht: Springer.
- Bjorkman, Brownwyn & Hedde Zeijlstra. 2019. Checking up on ( $\emptyset$ )-Agree. *Linguistic Inquiry* 50. 527–569.
- Chomsky, Noam. 2000. Minimalist inquiries : The framework. In Roger Martin, David Michaels & Juan Uriagereka (eds.), *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, 89–155. Cambridge, Mass. : MIT Press.
- Chomsky, Noam. 2001. Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale : A life in language*, 1–52. Cambridge, Mass.: MIT Press.
- Embick, David. 2015. *The morpheme. A theoretical introduction*. Boston: Walter de Gruyter.
- Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Ken Hale & Samuel Jay Keyser (eds.), *The view from Building 20. Essays in linguistics in honor of Sylvain Bromberger*, 111–176. Cambridge, Mass.: MIT Press.
- Pesetsky, David & Esther Torrego. 2004/2007. The syntax of valuation and the interpretability of features. In Simin Karimi, Vida Samiiian & Wendy K. Wilkins (eds.), *Phrasal and clausal architecture : Syntactic derivation and interpretation*, 262–294. Amsterdam: John Benjamins.

My most sincere thanks for the reviewers' comments.

---

# **A contribution of discourse analysis to the morphology of nominalizations. Study of the use of nominalizations in the genre of the scientific activity report using a corpus linguistics approach based on the Démonette and Lexique 3 databases.**

*Hugo Dumoulin*

MoDyCo, Université Paris Nanterre

---

The purpose of this paper is to explore the parameter of discourse genre in the study of nominalizations. How does discourse genre constrain the way nominalizations are used in discourse, from a morphological, syntactic and semantic point of view? We will look at these different aspects using a linguistic approach based on a corpus of professional writings in the activity report genre. Based on the work of Née, Sitri, Vénard (2016), who propose to make the link between genres and discursive routines, we propose to highlight the role of nominalizations in the specific phraseology of the activity report. In turn, it will be shown that discourse genre has affinities with certain types of verbal nominalizations, which helps to bring the discourse parameter to the fore in order to distinguish the competing suffix derivations -ion, -ment or -age, following the perspective developed by Missud & Villoing 2020.

## **1 Corpus and methodology**

### **1.1 Corpus**

The RapportS corpus was compiled as part of the ArchivU<sup>1</sup> project, which aims to approach institutional discourse from the angle of professional writings. It brings together the HCERES 2018 self-evaluation reports of 38 laboratories at the University of Paris Nanterre, which have been occluded and encoded in XML format<sup>2</sup>. To highlight the specific behaviour of nominalizations as a function of the discourse genre parameter, we rely on several comparison corpora: first of all, the Scientext corpus, which brings together texts of scientific articles (Tutin & Grossman 2014), secondly, the ADMIN corpus consisting of activity reports from two (non-scientific) French administrations (ADMR and Ucanss) in 2018, then several corpora from various discourse genre : the VCEUX corpus (Leblanc 2016), gathering the annual greetings of the presidents of the French republic since the 1970s, and the two corpora of the Lexique 3 database (New, Pallier, Brysbaert & Ferrand 2004), namely a written corpus composed of novels from the 1950-2000 period taken from Frantext<sup>3</sup>, and an corpus of film subtitles. The aim is to provide means of gaining a detailed understanding of the linguistic constraint exerted by the discourse genre, by drawing a comparison between genres that are far apart (argumentation/fiction) or more closely related by their theme (scientific article/scientific text) or function (activity report in the science field/in other fields).

### **1.2 Methodology**

Our tool-based linguistic approach is rooted in the general field of textometry (Lebart & Salem 1994). The TXM software (Heiden, Magué, Pincemin, 2010) is used for lemmatisation and automatic syntactic annotation (TreeTagger) of our corpus as a textual database. We chose to annotate very precisely the deverbal nominalizations in Xment, Xion, Xage of our corpora by using the Démonette database (Hatout & Namer 2014) via a Python script of our own. As a

---

<sup>1</sup> Labex *Les passés dans le présent*.

<sup>2</sup> The corpus represents 921,989 occurrences for 40,034 forms according to the segmentation and indexing performed by TXM.

<sup>3</sup> ATILF. *Base textuelle Frantext* (En ligne). ATILF-CNRS & Université de Lorraine. 1998-2023  
<https://www.frantext.fr/>

result, we carry out a set of lexicometric statistical observations, but also textometric ones, such as the calculation of specific co-occurrences.

## 2 Results

### 2.1. The distribution of nominalizations in RapportS compared with other corpora

#### 2.1.1. The contribution of verbal nominalizations in characterizing the genre of the report

We obtained initial lexicometric results (Figure 1): the frequency of deverbal nominalization tokens appears relatively high (3.40%) in RapportS and in ADMIN (2,93%) compared with VOEUX (0.96%), or even with LEXIQUE3 (0.36%). In addition, the type/token ratio appears very low in our corpus (0.058) compared with the VOEUX corpus (0.31), indicating a stronger fixation of vocabulary in the reports. In the light of these frequency calculations, verbal nominalizations appear to be somewhat characteristic of the genre of the professional activity report. Nevertheless, it appears that the frequency of deverbal nominalizations is also very high in the corpus SCIENTEXT of scientific texts (3,59%).

	Corpus size	Verbal nominalizations			
		Type	Token	Tokens frequency (%)	type/token ratio
RapportS	921 898	1 817	31 328	3,3982	0,0580
VOEUX	118 719	362	1139	0,9594	0,3178
SCIENTEXT	3 320 474		119396	3,5958	
ADMIN	43 145		1266	2,9343	
Lexique3_romans	~14 700 000			0,36089	
Lexique3_films	~50 000 000			0,20973	

Figure 1. Distribution of verbal nominalization among discourse genres

#### 2.1.2. An attraction of the -ion derivation for the discourse genre of the activity report and of the scientific article

A factorial correspondence analysis (Benzécri 1973) was carried out to represent the distribution of derivational suffixes among the corpora. The relationship between the modality of suffix and the modality of discourse genre appears to be statistically significant, although of low intensity ( $\phi = 0.1168$ ). It appears that the axis of greatest inertia opposes the suffix -ion to the other suffixes -ment and -age (87.83% of the variance), while the secondary axis opposes the suffixes -ment and -age to each other. Under these conditions, Figure 2 clearly shows a statistical attraction of the suffix -ion for the report (RapportS, Admin) and the scientific article (Scientext) genres, while -ment is significantly attracted by the Vœux and Lexique3\_romans corpora. Finally, -age appears to have a positive association with the Lexique3-films corpus. The more concrete nature of -age (Missud & Villoing 2021) seems to be confirmed by its attraction to the subtitle genre, representing oral interaction in written form. Conversely, we can link the preference for -ion in activity reports and scientific articles to the greater abstraction that characterizes it compared to -ment (Martin 2008: 165). Nevertheless, the -ion suffix does not clearly show a preference for the genre of the scientific article or the genre of the activity report.



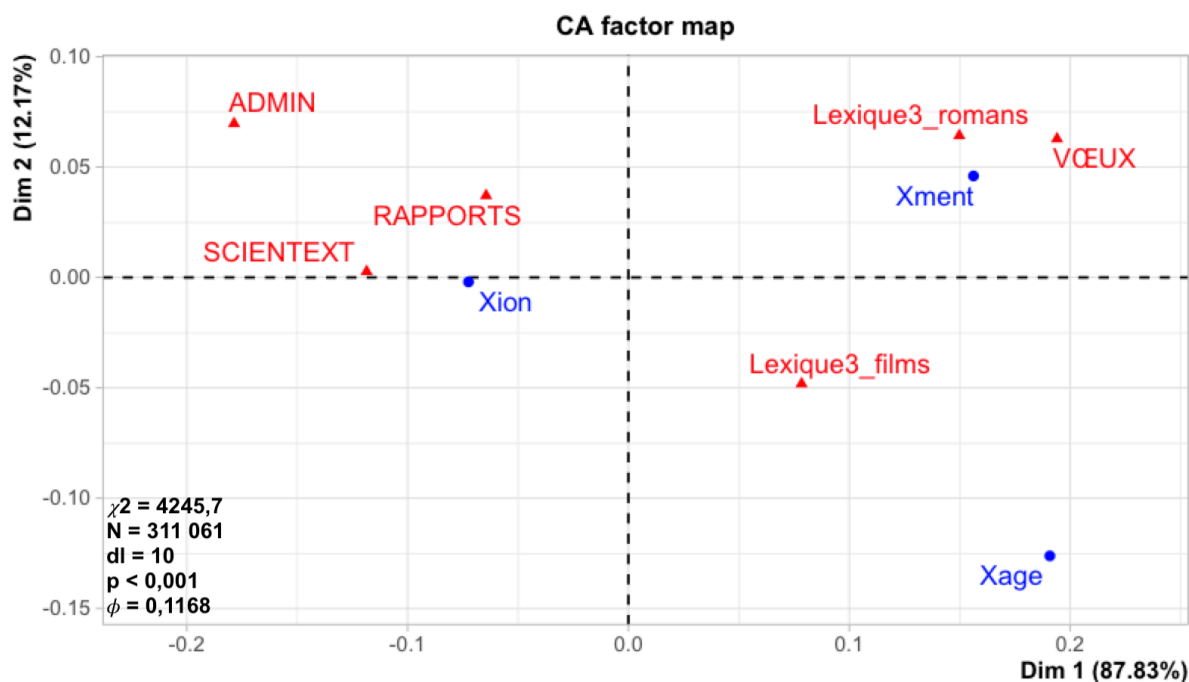


Figure 2. Correspondence factor analysis representing the distribution of derivational suffixes among the corpora

## 2.2. Textometric study of the corpus : nominalization as part of a routinization process

The textometric study now makes it possible to study the behaviour of nominalizations in the RapportS corpus by restoring them in their phrasal and textual context. In particular, we seek to identify the specific prepositional complements of nominalizations in the corpus, using Corpus Query Language (Figure 3). It appears that, among the statistically specific co-occurents (Lafon 1980), we find a large number of lexemes that are part of the lexical field of scientific activity: unit, research, knowledge, data, product, tool, ecosystem, knowledge, etc. It should be noted that the specific co-occurents of -ion ('unit', 'knowledge') are considerably more abstract than those of -ment ('tool', 'researcher'). But generally speaking, the use of nominalizations in the corpus is specifically linked to the representation of scientific activity - a central theme which carries the very function of the genre: we are dealing with a discursive routine characteristic of the genre which links deverbal nominalizations and the lexical field of scientific activity.

Requête [demonext="Xion"][frlemma="de du"]					Requête [demonext="Xment"][frlemma="de du"]				
Cooccurrent	Fréquence	CoFréquence	Indice	Distance moyenne	Cooccurrent	Fréquence	CoFréquence	Indice	Distance moyenne
NOM unité	1912	254	54	3.2	NOM recherche	5080	157	21	2.9
NOM produit	260	84	48	1.2	NOM outil	325	24	10	2.2
NOM connaissance	337	82	36	1.4	NOM chercheur	1200	48	10	2.3
VER:pres rechercher	56	35	32	3.0	NOM donnée	709	35	10	1.9
NOM savoir	332	76	32	2.0	ADJ nouveau	1440	53	10	2.5
NOM donnée	709	111	30	1.9	NOM intérêt	214	19	10	1.6
NOM écosystème	82	39	30	2.0	NOM fond	41	10	9	1.0
NOM recherche	5080	403	29	3.0	DET:ART un	13124	253	8	2.9
NAM Introduction	28	23	26	3.0	NAM CTEM	8	5	7	3.2
NOM colloque	1067	123	21	2.3	NAM signal	15	6	7	1.0
NOM activité	1364	143	20	3.3	NOM doctorants	922	35	7	2.4
					NOM professeur	351	20	7	2.5

Figure 3. Specific co-occurents in the prepositional position of nominalizations in -ion (left) and -ment (right). Distance to the right : 5 occurrences.

The importance of this result for the discourse semantics of nominalizations becomes apparent if we relate it to their syntactic properties. As Martin (2008) points out, -ion and -ment nominalizations are susceptible to multiple semantic subspecifications, leading to several readings: causative/inchoative, agentive/non-agentive. As a result, thematic roles that are clear for verbal complements become ambiguous for the prepositional complements of verbal nominalizations. For example, in the expression: "The development of new means of communication", there is ambiguity between the transitive and intransitive readings. Thus, the lexical field of scientific activity, which enters specifically into the position of complement of nominalizations, comes with an unclear semantic status. This contributes to the hypothesis of an abstract, routinised representation of research activity. The actants disappear, and things seem to happen by themselves - which may justify, notwithstanding a few remarks, the inclusion of nominalizations in the category of the "préconstruit" forged by the French school of discourse analysis (Pêcheux & alii 1979, Sériot 1986).

## Conclusion

Thus, both because of the types favoured by the genre of the activity report (abstract suffixes in -ion), and because of their textual role as co-occurents of the lexical field of scientific activity neutralising certain semantic values, verbal nominalizations can be associated with an effect of abstraction and discursive "routinization" specific to the activity report.

## Bibliography

- Benetti, L., & Corminboeuf, G., 2004, « Les nominalizations des prédicats d'action », *Cahiers de linguistique française*, 26, 413-435.
- Benzécri, J.-P., 1973, *L'analyse des données. Tome 1. La taxonomie*, Paris, Dunod, 675 p.
- Haas, P., Huyghe, R., Marin, R., 2008, « Du verbe au nom : calques et décalages aspectuels », *Actes du Congrès Mondial de Linguistique Française*, Paris, France, 2052-2065.
- Hathout, N. & Namer, F., 2014, "Démonette, a French derivational morpho-semantic network". *Linguistic Issues in Language Technology*, 11 (5), pp.125-168.
- Heiden S., Magué J.-Ph., Pincemin B., 2010, « TXM : une plateforme logicielle open-source pour la textométrie – conception et développement », 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010, Juin 2010, Rome, Italie, 1021-1032.
- Lafon P., 1980, « Sur la variabilité de la fréquence des formes dans un corpus », *Mots. Les langages du politique* 1, pp. 127-165.
- Lebart L., Salem A., 1994, *Statistique textuelle*. Dunod.
- Leblanc, J.-M., 2016, *Analyses lexicométriques des vœux présidentiels*, Londres, ISTE éditions.
- Martin, F., 2008, "The Semantic of Event Suffixes in French", in Schäfer, F. (ed.), *Working Papers of the SFB 732*, vol. 1., Stuttgart, University of Stuttgart.
- Missud, A., & Villoing, Fl., 2020, "The morphology of rival -ion, -age, and -ment selected verbal bases", *Lexique*, 26, pp. 29-52.
- Missud, A., & Villoing, Fl., 2021, "Investigating the distributional properties of rival -age suffixation and verb to noun conversion in French", *Verbum*, XLIII, pp. 41-68
- Née E., Oger C., Sitri F., 2017, « Le rapport : opérativité d'un genre hétérogène », *Mots* 114, Le rapport, entre description et recommandation, 9-24.
- Née E., Sitri F., Veniard M., 2016, « Les routines, une catégorie discursive pour catégoriser les genres ? », *Lidil* 53, 71-93.
- New, B., Pallier, Ch., Brysbaert, M. & Ferrand, L., 2004, "Lexique 2: A new French lexical database", *Behavior Research Methods, Instruments & Computers*, 36, 516-524.
- Pêcheux, M., Haroche, Cl., Henry, P., Poitou, J.-P., 1979, « Le rapport Mansholt : un cas d'ambiguïté idéologique », *Technologies, Idéologies, Pratiques*, 2, 1-83.
- Sériot, P., 1986, « Langue russe et discours politique soviétique : analyse des nominalizations », *Langages*, 81, 11-41.
- Tutin, A., & Grossman, F., 2014, *L'écrit scientifique : du lexique au discours ; autour de Scientext*, Presses universitaires de Rennes.

---

# Paradigmatic structures, defectivity, and the specificity of referents

Sophie Ellsäßer  
Universität Osnabrück

---

## 1 German indefinite pronouns

### 1.1 Paradigmatic structures

In German, there are different indefinite pronouns that can be used to refer to humans. Due to their development based on masculine nouns, each of them has very specific paradigmatic properties. Most cannot mark plurals (e.g. *einer* 'one'), some have exclusively masculine singular forms (e.g. *wer* 'who' or *jemand* 'someone'). Their paradigms are therefore often classified as defective in traditional approaches (cf. e.g. Pittner 1996 and Harnisch 2009 for *wer* and Thieroff 2011, 2012 for a contrasting view), with some showing more, some less possibilities for morphological differentiation.

### 1.2 Specificity of referents

These differences in paradigmatic properties have been linked to certain functional-semantic properties these pronouns had in older stages of German. For example, *einer*, which has masculine (*einer*), feminine (*eine*) and neutral (*eines*) singular forms, has been used for specific referents, while *jemand*, which has only masculine singular forms, was used for non-specific referents (cf. Fobbe 2004). In these contexts, the masculine indefinite pronoun *jemand* could refer generically to persons of any (or non-specified) gender (1) while the gender-specific forms of *einer* could refer to persons of specified masculine or feminine gender (2).

In contemporary German, however, these indefinite pronouns can be used for different degrees of specificity (cf. Haspelmath 1997). *Jemand*, for example, can also be found in contexts of specific reference (3). This increasingly dissolves the clear functional boundaries. When referring specifically to female referents, there are issues with the genericity of the masculine form *jemand*, which can be observed, for example, in varying agreement forms (3). For contexts like these, there are already some metalinguistic descriptions that classify them as being syntactically or pragmatically conspicuous (e.g. Kotthoff & Nübling 2018).

- |     |                                |            |                |                     |            |                |
|-----|--------------------------------|------------|----------------|---------------------|------------|----------------|
| (1) | <i>Kann</i>                    | <i>mir</i> | <i>bitte</i>   | <i>jemand</i>       |            | <i>helfen?</i> |
|     | can                            | me         | please         | someone.M.SG        |            | help           |
|     | 'Can someone help me, please?' |            |                |                     |            |                |
| (2) | <i>Da</i>                      | <i>ist</i> | <i>ein-e,</i>  | <i>die</i>          | <i>ich</i> | <i>kenne.</i>  |
|     | there                          | is         | one- F.SG      | who.F.SG            | I          | know           |
|     | 'There's one I know.'          |            |                |                     |            |                |
| (3) | <i>Da</i>                      | <i>ist</i> | <i>jemand,</i> | <i>der/?die</i>     | <i>ich</i> | <i>kenne.</i>  |
|     | there                          | is         | someone.M.SG   | who.M.SG /?who.F.SG | I          | know           |
|     | 'There's someone I know.'      |            |                |                     |            |                |

So far, there has been no comprehensive empirical study of the degrees of specificity of different indefinite pronouns in contemporary German or of the frequency or acceptability of these new functional-semantic contexts.

## 3 Theoretical framework

The special paradigmatic properties of indefinite pronouns as well as the functional-semantic properties linked to them offer an interesting testing ground for contemporary approaches to

defectiveness. Based on Sims (2015: 26), a defective paradigm can be defined as one that lacks a cell for a morphosyntactic or morphosemantic feature F, although this feature is defined for other representatives of this part of speech. Sims (2015) assumes that even though there are syntactic structures that would demand F, these structures would become ungrammatical when the corresponding lexeme from the defective paradigm was inserted.

On my poster, I will expand this concept from syntactic structures to reference semantic structures. By this definition, both indefinite pronouns would not be classified as defective in their original functional distribution. In older stages of German, they exhibited a strong degree of specialization in restricted contexts. *Jemand* prototypically occurred only in contexts where it refers generically to non-gender-specified referents via a masculine form. Thus, the feature FEMININE and PLURAL were not required. *Einer* referred to individual specific referents only. Therefore, the feature PLURAL was not required.

However, if *jemand* and *einer* do actually occur with new functional-semantic properties in contemporary German, this can affect their degree of defectivity according to the definition based on Sims (2015). This applies in particular for the feature FEMININE, for which the paradigm of *jemand* does not provide a distinct form, but which needs to be particularly expressed in contexts with reference to specifically female persons, as in (4).

(4) *Ich kenne jemanden, der/die dir helfen kann: Lea.*  
 I know someone.M.SG who.M.SG /?who.F.SG you help can Lea (name)  
 'I know someone who can help you: Lea.'

In these contexts, other features may be required which are not provided by the paradigms. In this case, the distinct feminine form of *jemand* is missing. The varying agreement forms of *jemand* as well as the metalinguistic discussion on the phenomenon, indicate that the masculine forms cannot necessarily be classified as grammatical in the corresponding contexts. The expansion of functional-semantic properties could thus be accompanied by an expansion of defectivity.

### 3 Empirical data

The poster is based on a contrastive analysis in the corpus *Mode- und Beauty-Blogs* 'fashion and beauty-blogs' which can be accessed via [dwds.de](http://dwds.de). The corpus contains informal written language data of contemporary German. Given the nature of the data, one can expect innovative grammatical structures.

In this corpus, evidence for *jemand* and *einer* is analysed, with forms of all genders referring to people. In order to analyse similar contexts and to gain information on varying agreement forms, only contexts with *jemand* or *einer/eine/eines* followed by a relative pronoun (as in (3) and (4)) have been examined. The degree of specificity is classified for each context. It is determined whether the indefinite pronouns refer to a specific or a non-specific referent. For specific referents, the social gender was identified in order to make statements about the need for the feature FEMININE. This allows us to determine the extent to which the original functional spectrums of *jemand* and of *einer* have expanded and provides an empirical base for the theoretical discussion on the degree of defectivity of the two indefinite pronouns.

The data indicate that the functional boundaries separating *einer* and *jemand* are increasingly dissolved in contemporary German: Both *einer* and *jemand* can be used with specific as well as non-specific reference. However, differences in the frequency of the functions are still evident. With *jemand*, non-specific reference is clearly more frequent than with *einer*. The empirical data show that the two indefinite pronouns undergo a process of functional expansion which is yet to be completed.

## 4 Aim of the poster

On the poster, I will compare the German indefinite pronouns *einer* and *jemand*. First, a brief overview concerning their paradigmatic properties will be given followed by a discussion of their defectivity based on previous and contemporary definitions of the notion. Afterwards, I will compare the degrees of specificity of *einer* and *jemand* based on the corpus analysis in order to discuss the interaction between defectivity and specificity in the data. I will also discuss indications of the word forms being classified as ungrammatical in certain contexts. The poster thus provides an insight into the interaction of paradigmatic structures and functional-semantic properties as the specificity of referents.

## References

- Fobbe, Eilika. 2004. *Die Indefinitpronomina des Deutschen: Aspekte ihrer Verwendung und ihrer historischen Entwicklung* (Germanistische Bibliothek 18). Heidelberg: Winter.
- Harnisch, Rüdiger. 2009. Genericity as a principle of paradigmatic and pragmatic economy. The case of German *wer* 'who'. In Walter Bisang, Hans H. Hock, Werner Winter, Patrick O. Steinkrüger & Manfred Krifka (eds.), *On Inflection* (Trends in Linguistics. Studies and Monographs [TiLSM]), 69–88. Berlin: De Gruyter.
- Haspelmath, Martin. 1997. *Indefinite pronouns* (Oxford studies in typology and linguistic theory). Oxford: Oxford University Press.
- Kotthoff, Helga & Damaris Nübling: *Genderlinguistik. Eine Einführung in Sprache, Gespräch und Geschlecht*. Tübingen: Narr.
- Pittner, Karin. 1996. Zur morphologischen Defektivität des Pronomens *wer*. In *Deutsch als Fremdsprache* (2). 73–77.
- Sims, Andrea D. 2015. *Inflectional defectiveness*. Cambridge: Cambridge University Press.
- Thieroff, Rolf. 2011. *Wer und was*. In *Germanistische Mitteilungen* 37(2). 47–64.
- Thieroff, Rolf. 2012. Die indeklinablen neutralen Indefinitpronomina. Etwas, was, irgendetwas, irgendwas und nichts. In Björn Rothstein (ed.), *Nicht-flektierende Wortarten* (Linguistik – Impulse und Tendenzen 47), 117–147. Berlin: De Gruyter.

---

# Morphology and spelling variation: A case study on handwritten German

*Stefan Hartmann*   *Kristian Berg*   *Daniel Claeser*  
HHU Düsseldorf   Universität Bonn   Universität Bonn

---

## 1 Introduction

Recent research has challenged some traditional assumptions of linguistic morphology. For instance, there is much evidence by now that morphological structure is a gradient, rather than categorical, phenomenon (Hay & Baayen, 2005) – for instance, psycholinguistic experiments have shown that a low-frequent item like *discernment* is more segmentable than high-frequent *government* (Hay, 2003, 136), indicating that morphological boundaries differ in their strength. Research on the interface of morphology and phonetics has shown that realizations of word-final *s* differ in subphonemic detail, conditional on its morphological status, e.g. whether it is a plural-*s*, a clitic *s*, or part of a stem (Schmitz, 2022). This indicates that morphological structure impacts the concrete realization of complex words in the spoken modality. But there is also some evidence that morphology has an effect on *written* language as well (e.g. Ernestus & Mak, 2005; Schmitz et al., 2018; Gahl & Plag, 2019; Chamalaun et al., 2021). In this paper, we follow up on previous research that has taken spelling errors as a starting point for investigating the relationship between morphology and writing. More specifically, we present a pilot study addressing the question of whether some morphological units are more susceptible to spelling errors than others. The hypothesis that this might be the case is motivated by the consideration that if the distinction between e.g. inflectional and derivational morphology is not just a matter of linguistic description but has correlates in one way or another in language users’ knowledge, we can expect units with a different morphological status to show processing- and production-related differences. In the present pilot study, we zoom in on one particular kind of spelling error that is arguably particularly informative about the morphology/graphemics interface, namely word-final one-letter and two-letter omissions in a dataset of handwritten exams.

## 2 Material and methods

We draw on the GraphVar corpus (Berg et al. 2021; see <https://graphvar.uni-bonn.de/> for details), a corpus of handwritten German A-level exams. It comprises 1,667 exams, covering the time span from 1923 to 2018. For the present, synchronically-oriented study, we use the 667 texts from 1990 onwards. The data have been transcribed, POS-tagged using the STTS tagset, lemmatized, and annotated with “target hypotheses” by trained annotators. The target hypothesis layer represents the orthographically correct form of each token according to the official orthography at the time when the exam in question was written.

For each combination of actual spelling and orthographically correct spelling, we semi-automatically filtered out the ones containing omissions, i.e. cases where the actual spelling contains fewer letters than the orthographically correct spelling. We further categorized these cases according to the position of the omission, and the constituent type containing it. We used a very broad classification into stems, inflectional and derivational affixes. We focus on nouns as the largest word category in the corpus (type-wise). We employed the annotation layers “IST” (actual appearance of the token in the exam) and “IST\_Ziel” (representing the target hypothesis) to automatically extract all tokens deviating from the expected (correct) form,

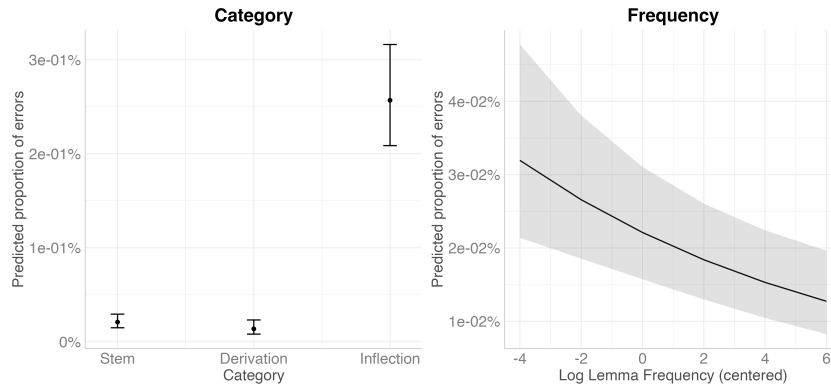


Figure 1: Effects plot for the two fixed effects in our model, “Category” and “Frequency”.

yielding more than 50,000 errors of different types. We discarded the largest group of errors, punctuation mistakes, as irrelevant for the given research question. We then limited the set of errors to nouns excluding proper nouns by filtering the remainder of observations for the STTS tag “NN”, yielding 14,200 instances of misspelled nouns. We discarded all capitalization errors and hyphenation errors in compounds and incorrectly separated detachable verb particles from this set. In a final filter step, we limited our analysis to tokens on the “IST” level containing fewer characters than expected as in the “NORMAL” annotation layer, thereby excluding the error types of insertion and replacement. The remaining set consisted of 919 nouns with initial, medial, and final character omissions. To keep the manual annotation work feasible, our manual analysis focused particularly on word-final omissions of individual letters and letter pairs. Two words were omitted from the analysis as they alternate between two different stem forms: *Friede/Frieden* ‘peace’ and *Glaube/Glauben* ‘faith’.

### 3 Results

#### 3.1 Distribution of omissions

As mentioned above, we will focus on word-final one- and two-letter omissions exclusively. There are 393 omissions of single word-final letters in our data, and 156 omissions of word-final bigrams. Table 1 shows the distribution of the single final letter omissions across the three morphological categories. For comparison, the rightmost column of Table 1 shows for all nouns in our dataset to which category the last letter of a word belongs.<sup>1</sup>

	Example actual spelling	Example correct spelling	Number of omission	Total # of occurrences
Stem	Grupp	Gruppe (‘group’)	46 (12%)	196,839 (46%)
Infl. suffix	Sprecher	Sprechers (‘speaker <sub>GEN</sub> ’)	331 (84%)	125,314 (30%)
Deriv. suffix	Wassertief	Wassertiefe (‘water depth’)	16 (4%)	102,030 (24%)

Table 1: Number of omissions of single final letters in nouns according to morphological type. Data base: Texts from the GraphVar corpus > 1990.

<sup>1</sup>The results were obtained by manually annotating the noun word form types in the corpus and then extrapolating to the token level.

We fit a mixed-effects model to the data, with “spelling error” (yes or no) as a binary response variable, morphological category (stem, inflectional suffix, derivational suffix) and lemma frequency (centered and log-transformed) as predictor variables, and random intercepts for the individual documents in which each word occurred. As Table 2 as well as the effects plot in Figure 1 show, both the morphological category of inflection and the lemma frequency of the word in question emerge as highly significant predictors. The former is in line with our hypothesis: Missing final letters in German nouns are much more likely on inflectional suffixes as opposed to stems or derivational suffixes. The latter can also be expected from a usage-based perspective as well as from a methodological point of view: Firstly, it makes intuitive sense that lexemes that occur less frequently are more error-prone than lexemes that language users encounter on a daily basis. Secondly, for low-frequency words, the proportion of spelling errors can easily be very high due to the comparatively small sum total of instances.

	Estimate	Std. Error	z value	Pr(>  z )
(Intercept)	-8.4	0.17	-49	< 2e-16 ***
CategoryDerivation	-0.42	0.29	-1.5	0.15
CategoryInflection	2.51	0.16	16.11	< 2e-16 ***
CenteredLogLemmaFreq	-0.092	0.025	-3.7	0.00021 ***

Table 2: Coefficients of the fixed effects in the regression model.

Turning to word-final bigrams, their omission is less frequent, but the pattern is even more striking, with 4 (3%) stem omissions, 146 (94%) inflectional suffix omissions, and 6 (4%) omissions of derivational suffixes. Importantly, we find no instance of an omission crossing a morphological boundary. If final bigrams were randomly left out, we would expect to find at least some cases like <Zustan> instead of <Zustands> (‘state-GEN’), where both the stem final <d> and the inflectional <s> are missing. Instead, omitted bigrams seem to respect morphological boundaries.

### 3.2 Morphological category

We also coded the spelling errors for the kinds of suffixes that the missing final letters are part of. Overall, as Table 3 shows, case marking is dominant in our data on final letter omissions. Omissions of the final letter of a case marking suffix are about four times as frequent as omissions of the final letter of a number marking suffix. Of course, the absolute numbers can be misleading without a comparison to non-omitted forms. As a baseline, we use morpho-syntactically annotated data from the Tüba/DZ corpus (Telljohann et al., 2017). A CoNLL-formatted treebank based on German newspaper texts of 1989 to 1999, Tüba/DZ comprises roughly 1.8 million tokens in about 3.600 newspaper texts. In the absence of a morphological layer in GraphVar, Tüba/DZ appears to be a reasonably comparable corpus with respect to domain and register to obtain distributions of case and number markings. To assess how many words that end with <-s> and <-n> there are (where this letter is not part of the lemma), and whether this letter is part of a case or a number marking, we make use of the morphological annotation layer in Tüba/DZ. The two rightmost columns in Table 3 show the results.

## 4 Conclusion

Our results suggest that missing final letters are not randomly distributed. They depend on the type of their morphological unit. This in turn indicates that until the very last stages of planning, this morphological information must be represented in the minds of writers – potentially in a



	-e (GraphVar)	-s (GraphVar)	-n (GraphVar)	total (GraphVar)	-s (Tüba/DZ)	-n (Tüba/DZ)
number	24	2 (1%)	39 (31%)	65 (22%)	2,491 (88%)	3,235 (81%)
case	-	151 (99%)	85 (69%)	236 (78%)	346 (12%)	776 (19%)

Table 3: Single final letter omissions of *-e*, *-s* and *-n* as part of inflectional suffixes, categorized according to the feature type that is marked (number vs. case).

“buffer” similar to the “articulatory buffer” in which Levelt (1989) assumes that the pieces of the “phonetic plan” are stored.

The overrepresentation of dedicated case as opposed to number affixes lines up with structuralist observations about the hierarchical ordering of inflectional categories. Eisenberg (2020, 162), for example, observes that number is the dominant and hierarchically higher category when compared with case. Summing up, then, our study provides further evidence for the key role of morphological units in the distribution of spelling errors, which in turn has far-reaching ramifications on theories of morphological storage and processing. While the pilot study presented here is certainly far from answering all pertinent questions, and will have to be complemented by follow-up studies going beyond word-final one- and two-letter omissions, we hope that it provides a valuable starting point for further in-depth corpus-based research on the interface between morphology and writing.

## References

- Berg, Kristian, Cedrek Neitzert & Jonas Romstadt. 2021. *Das Graphvar-Handbuch. Korpusaufbau und Annotationen*. work in preparation.
- Chamalaun, Robert J. P. M., Anna M. T. Bosman & Mirjam T. C. Ernestus. 2021. The role of grammar in spelling homophonous regular verbs. *Written Language & Literacy* 24(1). 38–80.
- Eisenberg, Peter. 2020. *Grundriss der deutschen Grammatik. Bd. 1: Das Wort*. Stuttgart: Metzler 5th edn.
- Ernestus, M. & W. M. Mak. 2005. Analogical effects in reading Dutch verb forms. *Memory and Cognition* 33. 1160–1173.
- Gahl, Susanne & Ingo Plag. 2019. Spelling errors in English derivational suffixes reflect morphological boundary strength: A case study. *The Mental Lexicon* 14(1). 1–36.
- Hay, J. B. & R. H. Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348.
- Hay, Jennifer. 2003. *Causes and consequences of word structure*. New York: Routledge. OCLC: ocm52485114.
- Schmitz, Dominic. 2022. *Production, perception, and comprehension of subphonemic detail. Word-final /s/ in English*. Berlin: Language Science Press.
- Schmitz, Tijn, Robert Chamalaun & Mirjam Ernestus. 2018. The Dutch verb-spelling paradox in social media: A corpus study. *Linguistics in the Netherlands* 35(1). 111–124.
- Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister & Kathrin Beck. 2017. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Tübingen: University of Tübingen. <https://sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1707.pdf>.

---

# Base ellipsis in coordinative constructions: the case of *pré et post-X* in French

Kentaro Koga  
Fukuoka University

---

## 1 Introduction<sup>1</sup>

Coordination is a syntactic phenomenon where two or more elements, sharing the same syntactic category, are juxtaposed (Goodall 2017). The conjuncts (i.e. juxtaposed elements) are free morphemes or phrases, whereas bound morphemes such as prefixes cannot be conjuncts by themselves.

However, in French, we can observe the cases where two prefixes are apparently linked by the conjunctions *et* ‘and’ or *ou* ‘or’,<sup>2</sup> such as example (1)<sup>3</sup>:

- (1) *les garderies pré et post-scolaires*  
ART.PL nursery.PL pre and post-scholastic.PL  
‘before and after-school childcare’

How can we account for the morphosyntactic status of the first prefix (i.e. *pré*)? Is it really possible that two bound morphemes are directly conjoined?

Indeed, certain prefixes can appear as morphologically independent elements, but this is not the case for the French prefix *pré*. In fact, in the case of simple affixation, the attachment to a base lexeme is necessary for this prefix (cf. 2a), and the base ellipsis is not possible (2b):

- (2) a. *la préinscription est gratuite.*  
ART.F.SG pre-registration.F.SG be-3SG free.F.SG  
‘the pre-registration is free’  
b. *\*la pré est gratuite.*

From these facts, we can suppose that the prefix *pré* does not have the status of an autonomous lexical item, and that the ellipsis of the base is exceptionally possible for coordinative constructions where the two prefixes share the same base. We will call the units such as *pré et post-scolaires* (cf. 1) “*pré et post-X* constructions”. However, this does not imply that the constructions do not accept other pairings of prefixes.

In this study, we will focus on constructions where the first prefix is *pré*. Firstly, we will examine the difference between lexicalized prefixes and the prefix *pré* in the *pré et post-X* constructions. Secondly, we will analyze the frequency and the possibility of combinations of prefixes (*pré* and other prefixes) by means of the frTenTen20 corpus, which consists of 20.9 billion words collected from the web. Based on the result of the corpus analysis, we will argue that the base ellipsis is not a simple morphological process but a process to be realized within syntactic operations.

---

<sup>1</sup> This research was supported by JSPS KAKENHI (Grant Number 23K12184) and funding from Fukuoka University (Grant Number GW2302).

<sup>2</sup> Other conjunctions, such as *ni* ‘nor’ can also be used, but far less frequently than *et* and *ou*. In this research, we will focus on the cases with *et* and *ou*.

<sup>3</sup> All examples from (1) to (7) are from the texts registered in frTenTen20 corpus.

## 2 *Pré*: non-autonomy and non-lexicalization

There are two kinds of prefixes in French, namely, autonomous (or separable) prefixes and bound prefixes (Lehmann and Martin-Berthet 1998: 148). The former can function not only as a prefix, but also as a preposition or an adverb. For example, *après* in *après-guerre* ‘postwar’ is a prefix attached to the base *guerre* ‘war’, whereas in *après la guerre* ‘after the war’, it is a preposition (Amiot 2005: 183-184). On the other hand, bound prefixes cannot be morphologically autonomous elements (cf. 2b). Both *pré* and *post* are bound prefixes.

Certain bound prefixes can appear alone as an ellipped form of a complex lexical unit. In this case, it is possible that the prefix is lexicalized, containing semantic information that the base originally had.

For example, the prefix *ex* can appear alone if the ellipped base denotes a romantic or spousal partner (e.g. *mon ex-copain* ‘my former boyfriend; *mon ex-femme* ‘my former wife’). In contrast, *mon ex-conseiller* ‘my former counselor’ cannot be transformed into *mon ex*. This restriction suggests that the autonomous *ex* already contains the meaning of its ellipped base (i.e. that of *copain*, *femme*, etc.). In effect, we can observe the speech such as *vous pensez à votre ex ?* ‘do you think of your former (boyfriend, husband, etc.)?’ without specifying which is the base lexeme exactly. Given these facts, we can consider that the autonomous *ex* is lexicalized, having a meaning “former romantic or spousal partner”.<sup>4</sup>

On the other hand, *pré* in the *pré et post-X* construction is not a lexicalized prefix because, unlike the autonomous *ex*, the indication of the base is always necessary. It is always the first apparition of the base (i.e. *X* of *pré-X*) that can be ellipped (cf. 3a), and not vice versa (cf. 3b):

- (3) a. *les pré et post-événements*  
 ART.M.PL pre and post-events.M.PL  
 ‘pre(-event) and post-event celebrations’  
 b. *\*les pré-événements et post*

The base of the *pré et post-X* construction is a noun (cf. 3a) or a relational adjective (i.e. an adjective having a classificatory function in relation to the head). It is often the case that a nominal lexeme is applied without suffixation, instead of the corresponding adjective that is formally absent (cf. 4):<sup>5</sup>

- (4) *en période pré et post Covid-19.*  
 in period.F.SG pre and post Covid-19.SG  
 ‘in before and after-Covid-19 period’

It is the coordinative construction that supports the functional compatibility of the conjuncts. Since the prefix *pré* is not lexicalized, the conjoined elements are therefore *pré-X* and *post-X*. In other words, without the support of coordinative construction, the ellipsis of the base of *pré-X* becomes impossible (e.g. *\*en période pré*; cf. *en période pré Covid-19*).

<sup>4</sup> Similar phenomena can be observed for clipped lexemes. According to Kerleroux (1999:97-102), *colo* (the clipped form of *colonie* ‘colony’), signifies only *colonie de vacances* ‘summer camp’.

<sup>5</sup> As is the case in (4), the hyphen between *post* and the base can be replaced by a space. In some cases, there is no hyphen or space between the two (e.g. *la période pré et postélectorale* ‘before and after the election period’).

### 3 Base ellipsis: a process activated within syntactic operations

We can observe 5,488 occurrences of “*pré et*” sequences and 1,055 occurrences of “*pré ou*” sequences in the frTenTen20 corpus. As Table 1 indicates, the most frequent element co-occurring with *pré* is the prefix *post*, with 4,927 occurrences of *pré et post-X* and 851 occurrences of *pré ou post-X*:

**Table 1:** Co-occurrence of “*pré et/ou*” (frTenTen20 corpus)

(a) “ <i>pré et</i> ”		(b) “ <i>pré ou</i> ”	
Sequence	Occurrence	Sequence	Occurrence
1 <i>pré et POST</i>	4,927	1 <i>pré ou POST</i>	851
2 <i>pré et PRO</i>	155	2 <i>pré ou PROTO</i>	46
3 <i>pré et PROTO</i>	148	3 <i>pré ou PÉRI</i>	33
4 <i>pré et PARA</i>	79	4 <i>pré ou PER</i>	21
5 <i>pré et PÉRI</i>	66	5 <i>pré ou RÉTRO</i>	20
6 <i>pré et PER</i>	45	6 <i>pré ou PRO</i>	18
7 <i>pré et INTRA</i>	17	7 <i>pré ou PARA</i>	8
8 <i>pré et RÉTRO</i>	10	8 <i>pré ou NÉO</i>	8
9 <i>pré et EXTRA</i>	9	9 <i>pré ou EXTRA</i>	6
others	32	others	44
<b>TOTAL</b>	<b>5,488</b>	<b>TOTAL</b>	<b>1,055</b>

The high frequency of *pré et/ou post-X* can be described in terms of the semantic compatibility of the two prefixes, anteriority and posteriority, respectively. In addition to this, there are also some occurrences of *pré et/ou après-X*, where the autonomous prefix *après* ‘after’ indicates posteriority.

The base lexeme is varied in general, but in some sequences, the pairing of prefixes and their bases is almost fixed. For example, in the sequence of *pré et/ou proto-X*, the item coming after is, 89.2% of the time, *histoire* ‘history’ or lexemes derived from this noun (*historique* ‘historic’, *historien* ‘historian’, etc.).

In many cases, the coordination occurs within the word level, such as with the relational adjectives *pré-natal* and *post-natal* in (5):

(5) *un atelier sur le yoga pré et post-natal*  
 ART.M.SG workshop.M.SG on ART.M.SG yoga.M.SG pre and postnatal.M.SG  
 ‘a workshop on pre(natal) and postnatal yoga’

However, in the corpus, there are around 200 cases (approximately 3% of the total occurrences) containing other (especially, functional) elements such as prepositions and articles. In (6), two prepositional phrases modifying the head *responsable* are conjoined. The coordination therefore occurs on a syntactic level:

(6) *Responsable de la pré et de la postproduction*  
 responsible.M.SG of ART.F.SG pre and of ART.F.SG post-production.F.SG  
 ‘person responsible for pre(production) and postproduction’

The example in (6) demonstrates that the ellipsis of the base of *pré* is not a simple morphological process, but a process activated within syntactic operations.<sup>6</sup> In addition, *pré et post-X* construction is not a completely fixed unit. The process of coordination embedded in this construction accepts syntactic units such as prepositional phrases. In this sense, this construction is different from fixed *X-et-Y* sequences such as *Lot-et-Garonne* (the name of a department in France, consisting of the names of two rivers).

Furthermore, the elements conjoined with *pré-X* may not only be prefixed elements but also elements with a modifier of a noun (cf. 7):

- (7) a. *commentaires de pré et début de match*  
 comments.M.PL of pre and beginning.M.SG of match.M.SG  
 ‘comments of before and the beginning of match’
- b. *aux pré et jeunes adolescents*  
 to.ART.PL pre and young.M.PL adolescents.M.PL  
 ‘to pre(adolescents) and young adolescents’

In the examples (7a) and (7b), the categorical difference between the (bound) prefixes and the modifier of noun is neutralized. Since the functions of prefixes and modifiers are identical (i.e. to modify the base or head), they can be conjoined together in *pré et post-X* construction.

## 4 Conclusion

The base ellipsis in *pré-X* is a process that is not available alone. This process needs to be activated within *pré et post-X* construction. From the observation of the corpus, it is evident that the pairings of the conjuncts are not limited to a specific prefix but are open to other types of modifiers. This fact shows that a process of prefixation and a syntactic process may coexist in the same construction.

## References

- Amiot, Dany. 2005. Between compounding and derivation: elements of word formation corresponding to prepositions. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer and Franz Rainer (eds.), *Morphology and its demarcations*, 183-195. Amsterdam: John Benjamins.
- Goodall, Grant. 2017. Coordination in Syntax. In *Oxford Research Encyclopedia of Linguistics*. <<https://doi.org/10.1093/acrefore/9780199384655.013.36>>. Published online by Oxford University Press, 29 March 2017.
- Lehmann, Alise. & Martin-Berthet, Françoise. 1998. *Introduction à la lexicologie : sémantique et morphologie*. 3<sup>rd</sup> edition (2010). Paris: Armand Colin.
- Kerleroux, Françoise. 1999. Sur quelles bases opère l’apocope ? In *Sillexicales 2*, 95-106.
- Meinschaefer, Judith. 2023. Morphological ellipsis in coordination in Romance. In Natascha Pomino, Eva-Maria Remberger and Julia Zwink (eds.), *From formal linguistic theory to the art of historical editions: the multifaceted dimensions of Romance linguistics*, 49-66. Göttingen: V&R unipress.

---

<sup>6</sup> The postlexical feature of the ellipsis can also be accounted for from a phonological point of view. Meinschaefer (2023) reports that the possibility of ellipsis in coordination may depend on the number of syllables that the base lexeme has.

---

# Gender agreement in Italian compounds with *capo-*

Irene Lami<sup>1</sup>, M. Silvia Micheli<sup>2</sup>, Jan Radimský<sup>3</sup> & Joost van de Weijer<sup>1</sup>

<sup>1</sup>University of Lund, <sup>2</sup>University of Milano – Bicocca, <sup>3</sup>University of South Bohemia

---

## 1 Introduction

This study aims to investigate the gender and number inflection of a particularly productive compound type in Italian, namely the Noun Noun compounds made up of the word *capo* (literally 'head', see below on semantics) as the left-hand constituent and head of the compound (e.g., *capostazione*, chief.station 'station master'). Based on an experimental test, we will provide an overview on the strategies that native speakers employ in gender and number inflection, focusing on the formation of the feminine.

The history of this type of compound already begins in Latin, but it is only in more recent stages of the language that this pattern acquires remarkable productivity. As shown in the diachronic analysis by Micheli (2020: 130-140), already in Medieval Latin, the word *caput* occurred together with another noun in combinations where it referred to 'beginning' (as in *caput anni* 'early part of the year') but more frequently 'chief, person in a leading position' (as in *caput castris* 'chief of the military camp'). While in Old Italian *capo-* occurs within compounds expressing a wide range of meanings (e.g., 'head', 'initial part', 'leader'), from the 16th and 17th centuries onwards, it specialises in the creation of agentive compound nouns, indicating persons in positions of leadership or power (Micheli 2020: 149).

In present-day Italian, agentive compounds with *capo-* represent a well-established pattern, which includes a not entirely homogeneous set of words. Following the classification proposed by Bisetto & Scalise (2005), within this category we can identify compounds in which *capo-* is linked by a subordinative relationship to a second element that is a noun referring to a place, an institution, or a human group (e.g., *caposquadra*, chief.team 'foreman') and compounds where the second constituent is an agentive noun (e.g., *caporedattore*, chief.editor 'editor-in-chief'). The relationship between the two constituents in the latter type of compounds can be interpreted in three ways: i) as a coordinative relationship (i.e., *caporedattore* is the one who is both the editor and the chief); ii) as a subordinative relationship (i.e., *caporedattore* refers to the chief of the editors); iii) as an attributive relationship (i.e., *capo* represents the modifier of the noun *redattore*, which represents the head of the compound).

However, the interest in this type of compounds is not only a matter of semantics: indeed, they represent a highly interesting category for observing inflectional phenomena within Italian compound words. As shown by Micheli (2016: 25-28), the number inflection of compounds with *capo-* displays numerous instances of overabundance (i.e., «the situation in which two or more inflectional forms are available to realize the same cell in an inflectional paradigm» according to Thornton 2019), especially within the subtype where the second constituent is an agent noun (e.g., the plural form of *caporedattore* 'editor-in-chief' can be *capipLredattoreSG*, *capoSGredattoriPL*, and *capipLredattoriPL*).

In contrast, the gender inflection of compounds with *capo-* has not been systematically investigated so far. The issue of gender in this type of compound words is relevant for two reasons: on the one hand, the presence of more than one plural form suggests that gender may also be an irregular phenomenon; on the other hand, the gender inflection of the word *capo* as free form appears problematic, as it is generally considered by dictionaries to be a masculine noun (based on its use with the meaning 'head') which should not be inflected according to feminine gender.

Moreover, compounds with *capo-* fall into the category of occupational titles, which in Italian have been the matter of debate (both among specialists and in the general public) from a

sociolinguistic point of view. More specifically, since the seminal works by Sabatini (1985; 1986), it has been observed that they represent a crucial aspect of the use of sexist language, being many nouns indicating prestigious professions traditionally performed by men (e.g., *ministro*<sub>M</sub> ‘minister’) still often inflected only in the masculine form even when referring to a female subject (see, among others, Gheno 2019; Thornton 2012; Zarra 2017). Although this issue also concerns occupational nouns expressed through a compound with *capo-*, these forms are often neglected. Interestingly, when they are mentioned in the guidelines advocating a non-sexist use of language, they are treated as invariable nouns, where therefore the feminine form should not be used (e.g., Telve 2011; Robustelli 2012; Gheno, 2018).

Our study aims at filling this gap and focusing on the gender inflection of *capo-* compounds. Particularly, this research aims to answer the following questions: are there differences in the gender inflection between compounds with *capo-* and other occupational titles? What strategies do speakers implement when they decide to inflect for gender a compound with *capo-*? Is there a difference between the inflection of *capo-* in isolation and *capo-* within a compound? We will investigate the agreement strategies adopted by speakers and correlate them with both morphological and sociolinguistic parameters.

## 2 Methods

The test was administered through the Sogolytics platform. Informants should be asked to listen to ten sentences with masculine referents, and then to inflect the whole sentence asking to change the referent(s) from a masculine one to a feminine one, both in singular and plural. The following five categories of target nouns were investigated: 1) occupational titles traditionally linked to women; 2) occupational titles traditionally linked to men; 3) *capo* in isolation; 4) subordinate compounds with *capo-*; 5) compounds with *capo-* and an agentive noun. Each category was represented by eight nouns, shown in Table 1. The occupational titles traditionally linked to women were selected based on their acknowledged ‘unproblematicity’ regarding feminine inflection in the literature (see among many others Sabatini, 1987; Proudfoot & Cardo, 2005; Cortelazzo, 2017; Giusti & Iannàccaro, 2020; Ricci 2021; Sulis & Gheno, 2022). In the selection, we aimed to find words similar in frequency when inflected for feminine. The occupational titles traditionally linked to men were selected on the basis of their acknowledged resistance of accepting a feminine inflection. These words were previously pointed out by Zarra (2017) in his analysis of gender inflection for titles and professions to which women historically have had limited access (see also Miglietta, 2022). Professional titles with a possible double ending (e.g., *avvocatessa*, *avvocata*, lawyer.F) were not included in this category because of theoretical considerations and the fact that experimental studies have suggested that a gender bias is linked to the traditional affix *-essa* but not to the modern one *-a* (Mucchi Faina & Baino 2006; Merkel et al. 2012; Merkel 2013). The compounds containing *capo-* were extracted from the Zingarelli dictionary (online version, 2022).

Traditionally women	Traditionally men	Subordinate compound	Coordinate compound
<i>infermiere</i> “nurse”	<i>chirurgo</i> “surgeon”	<i>capotreno</i> “train conductor”	<i>caporedattore</i> “editor in chief”
<i>maestro</i> “teacher”	<i>architetto</i> “architect”	<i>caporeparto</i> “department head”	<i>capocarceriere</i> “head of prison”
<i>commesso</i> “shop assistant”	<i>sindaco</i> “mayor”	<i>capoclasse</i> “class monitor”	<i>capocameriere</i> “headwaiter”
<i>cassiere</i> “cashier”	<i>deputato</i> “member of parliament”	<i>capogruppo</i> “group leader”	<i>capocronista</i> “news editor”
<i>segretario</i> “secretary”	<i>ministro</i> “minister”	<i>capogabinetto</i> “head of the cabinet”	<i>capomaestro</i> “master builder”
<i>ballerino</i> “dancer”	<i>magistrato</i> “magistrate”	<i>capobranco</i> “pack leader”	<i>capocomico</i> “lead comic”

<i>portiere</i> “conciierge” <i>cameriere</i> “waitress”	<i>ingegnere</i> “engineer” <i>assessore</i> “assessor”	<i>caposezione</i> “head of the section” <i>caposquadra</i> “foreman”	<i>capocuoco</i> “chef” <i>capooperaio</i> “head laborer”
---	--	--	--

**Table 1:** Target nouns

### 3 Preliminary results

The test was completed by 192 respondents. Of these, 134 were women, 57 were men, and 1 did not identify as either male or female. All were native speakers of Italian, but six of them reported a second native language (i.e., Croatian, Sardinian, English, Spanish, Turkish or German). Their age ranged from 19 to 74 years, with an average of approximately 44 years. 168 respondents lived in Italy at the time of the survey, the remaining 24 reported living in another country. The highest educational level obtained by 32 participants was middle school or high-school, that of the remaining 160 respondents had obtained some form of university degree. The participants’ responses towards gender-fair language were generally favourable. More than 97% of them indicated that they had at least some awareness of gender-fair language, more than 90% used it at least every now and then, and more than 85% had a neutral or a positive attitude towards it. The attitude towards gender-fair language correlated weakly with participant age (the younger participants had a somewhat more favourable attitude towards gender-fair language than the older ones), with participant gender (female participants’ had a somewhat more favourable attitude towards gender-fair language male participants) and academic degree (attitude towards gender-fair language was somewhat more favourable in participants with a university degree than in those without it). Table 2 shows the distribution of the grammatical categories of the nouns.

	Traditionally women		Traditionally men		Isolation		Subord. comp.		Coord. comp.	
	Sing	Plur	Sing	Plur	Sing	Plur	Sing	Plur	Sing	Plur
Feminine	0.97	0.92	0.86	0.83	0.62	0.69	0.10	0.18	0.06	0.31
Masculine	0.02	0.03	0.14	0.17	0.33	0.26	0.84	0.81	0.94	0.68
Other	0.01	0.06	0.00	0.00	0.05	0.05	0.06	0.01	0.00	0.02

**Table 2:** Preliminary results

The table shows that the proportions of feminine responses varied considerably across the five compound types. Within the first three types, the feminine responses dominate, while in the last two types the masculine responses dominate. In addition, the numbers suggest that the effect of compound number is not the same for each compound type. The proportion of feminine responses is *lower* in the plural forms than in the singular forms for female and male dominated professions, and *higher* in the plural forms for the remaining three compound types.

Starting from this data, in the presentation we will discuss in more detail the strategies adopted by speakers and offer some reflections on the relationship between number inflection and gender.

### References

- Bisetto, Antonietta & Sergio Scalise. 2005. The classification of compounds. *Lingue e Linguaggio* 4(2). 319-332
- Cortelazzo, Michele. 2017. “Il presidente, la presidente, la capra”. Web blogpost, Michele Cortelazzo *Parole. Opinioni, riflessioni, dati sulla lingua*. <https://cortmic.myblog.it/presidente/>.
- Gheno, Vera. 2018. Tutti i modi dell’hate speech sui social media: quando la lingua separa e ferisce. *Agenda Digitale*, 3.5.2018.
- Gheno, Vera. 2019. *Femminili singolari. Il femminismo è nelle parole*. Firenze: Effequ.



- Giusti, Giuliana & Gabriele Iannàccaro. 2020. Can gender-fair language combat gendered hate speech? Some reflections on language, gender and hate Speech. In Giuliana Giusti & Gabriele Iannàccaro (eds.), *Language, Gender and Hate Speech. A Multidisciplinary Approach*, 9-20. Venezia: Edizioni Ca' Foscari - Digital Publishing.
- Merkel, Elisa. 2013. *The two faces of gender-fair language*. Doctoral thesis. University of Padua.
- Merkel, E.; Maass, Anne, & Laura Frommelt. 2012. Shielding women against status loss: The masculine form and its alternatives in the Italian language. *Journal of Language and Social Psychology* 31(3). 311-320.
- Micheli, M. Silvia. 2016. Limiti e potenzialità dell'uso di dati empirici in lessicografia. Il caso del plurale delle parole composte. *Ricognizioni* 6(2). 15-33.
- Micheli, M. Silvia. 2020. *Composizione italiana in diacronia. Le parole composte dell'italiano nel quadro della Morfologia delle costruzioni*. Berlin/New York: De Gruyter.
- Miglietta, Annarita. 2022. Morfologia diacronica e parità di genere. *Italiano LinguaDue* 14(1). 861-878.
- Mucchi Faina, Angelica, & Martina Barro. 2006. Il caso di "professoressa": espressioni marcate per genere e persuasione. *Psicologia sociale* 1(3). 517-530.
- Proudfoot, Anna, & Francesco Cardo. 2005. *Modern Italian Grammar: A practical guide*. Second edition. New York, NY: Routledge.
- Ricci, Sara. 2021. *Stereotypes, prestige and grammar: occupational job titles in Italian*. Master Thesis. Ca' Foscari University of Venice.
- Robustelli, Cecilia. 2012. Linee guida per l'uso del «genere» nel linguaggio amministrativo. Progetto genere e linguaggio. Parole e immagini della comunicazione, svolto in collaborazione con l'Accademia della Crusca, Comune di Firenze, Comitato Pari Opportunità: [www.accademiadellacrusca.it](http://www.accademiadellacrusca.it).
- Sabatini, Alma 1985. Occupational Titles in Italian: Changing the Sexist Usage. In M. Hellinger, (ed.) *Sprachwandel und feministische Sprachpolitik: Internationale Perspektiven*. VS Verlag für Sozialwissenschaften.
- Sabatini, Alma. 1986. *Raccomandazioni per un uso non sessista della lingua italiana. Per la scuola e l'editoria scolastica*. Roma: Presidenza del Consiglio dei Ministri.
- Sabatini, Alma. 1987. *Il sessismo nella lingua italiana*. Commissione nazionale per la realizzazione della parità tra uomo e donna. Presidenza del Consiglio dei Ministri. Roma: Istituto Poligrafico e Zecca dello Stato.
- Sulis, Gigliola & Vera Gheno. 2022. The Debate on Language and Gender in Italy, from the Visibility of Women to Inclusive Language (1980s–2020s), *The Italianist* 42(1). 153-183.
- Telve, Stefano. 2011. Maschile e femminile nei nomi di professione. In *Enciclopedia Treccani*: [https://www.treccani.it/enciclopedia/maschile-e-femminile-nei-nomi-di-professione-prontuario\\_%28Enciclopedia-dell%27Italiano%29/](https://www.treccani.it/enciclopedia/maschile-e-femminile-nei-nomi-di-professione-prontuario_%28Enciclopedia-dell%27Italiano%29/).
- Thornton, Anna M. 2012. Quando parlare delle donne è un problema. In Anna M. Thornton & Miriam Voghera (eds.), *Per Tullio De Mauro. Studi offerti dalle allieve in occasione del suo 80° compleanno*, 301-316. Roma: Aracne.
- Thornton, Anna M. 2019. Overabundance: a canonical typology. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds.), *Competition in inflection and word-formation*. Cham: Springer.
- Zarra, Giuseppe. 2017. I titoli di professioni e cariche pubbliche esercitate da donne in Italia e all'estero. In Yorick Gomez Gane (ed.), «*Quasi una rivoluzione*». *I femminili di professioni e cariche in Italia e all'estero*, 19-104. Florence: Accademia della Crusca.
- Zingarelli, Nicola. 2022. *Lo Zingarelli 2022. Vocabolario della lingua italiana*. Bologna: Zanichelli.

---

# PrinParLat: a resource of Latin principal parts

Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, Giovanni Moretti

CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milano

---

## 1 Introduction

In this talk, we present PrinParLat, a free lexical resource documenting Latin verb inflection, making use of notions of theoretical morphology to provide rich information in a compact way.

Firstly, PrinParLat is a collection of principal parts. This notion was already used in traditional Latin dictionaries, where for each entry the citation form is accompanied by a set of forms from which the full paradigm can be inferred – e.g., the present active infinitive *amāre*, the first-person singular perfect active indicative *amāvī* and the perfect participle *amātum* for AMŌ ‘love’ in the Oxford Latin Dictionary – and it has recently been implemented in a principled fashion in theoretically grounded studies that investigate the implicative structure of paradigms with different approaches (Stump & Finkel, 2013; Bonami & Beniamine, 2016).

Secondly, two different layers of lexical units are used: each principal part is linked not only to the corresponding lexeme, but also to the corresponding flexeme(s). This distinction was introduced by Fradin & Kerleroux (2003) to account for cases of lexical items with different meanings but with the same form in all paradigm cells – e.g., for the French noun FILLE, there are two different lexemes (one for the meaning ‘girl’, one for the meaning ‘daughter’) that map to the same flexeme, as the wordforms are the same – and it has been recently applied (Bonami & Crysmann, 2018; Thornton, 2018) to the converse case of lexical items with the same meaning but different forms – i.e., to cases of overabundance; e.g., for the Italian noun ORECCHIO/A there are two different flexemes (one for the masculine forms *orecchio* SG and *orecchi* PL, one for the feminine forms *orecchia* SG and *orecchie* PL) that map to the same lexeme, as the meaning is the same (‘ear’).

Thirdly, regarding the inflectional behaviour of lexical items, information on the traditional conjugations of Latin verbs is provided. These can be considered as inflection “macro-classes” (Dressler, 2002; Beniamine et al., 2017), as they group items that are inflected similarly, but not identically – namely, they are inflected in the same way in imperfective wordforms, but not in the other ones. Furthermore, lexical items are also classified according to their fine-grained inflection “micro-class”, grouping together the ones that are inflected in the same way across the whole paradigm. These micro-classes are identified in an abstractive fashion (Blevins, 2016), by inspecting the alternation patterns that occur between all possible combinations of the listed principal parts.

## 2 The resource

The data of the resource are taken from the database of a morphological analyzer of Latin, Lemlat (Passarotti et al., 2017). The stems and endings reported there for verbs have been used to generate the full wordforms that we choose as principal parts. To restrict the remarkable time span covered by the Latin language, we only select about 8,000 entries that come from dictionaries of Classical Latin, thus excluding Medieval Latin verbs recorded in the database.

The resource is structured as a relational database, using the tables and columns defined in an emerging standard format for paradigmatic lexicons, Paralex. The core part is the forms table (1a), where for each principal part, we provide information on the cell it fills, the lexeme

it belongs to, and its form. Due to the unclear epistemological status of the actual pronunciation of Classical Latin, that can only be reconstructed, we provide orthographic transcriptions, rather than phonetic/phonological ones. We follow the traditional usage of Latin grammars and dictionaries in selecting PRF.ACT.IND.1.SG and PRF.PTCP.NOM.N.SG as principal parts from which perfective wordforms and forms displaying Aronoff (1994)'s Third Stem can be inferred, respectively (e.g., *amāvī*; *amātum*, for the verb meaning 'love'). We depart from the tradition in selecting PRS.ACT.INF and FUT.ACT.IND.3.SG – rather than the citation form PRS.ACT.IND.1.SG – as the principal parts from which imperfective wordforms can be inferred. This is due to the fact that the first-person singular is actually poorly informative on the content of other cells, as it neutralizes the opposition between 1<sup>st</sup> and 3<sup>rd</sup> conjugation verbs. Furthermore, an additional principal part is provided, namely FUT.PTCP.NOM.N.SG, to be able to infer future participle forms also in the few cases in which they display a stem different than the one of perfect participle forms. Additional principal parts are also needed for defective lexemes: for instance, we use the corresponding passive forms for deponent verbs that lack the active ones.

Additional information is provided in separate tables. For instance, regarding cells, we rely on the traditional description of the Latin verbal system, as documented in the features-values table (1c). However, in the cells table (1b), the corresponding notation in the UniMorph format is given (McCarthy et al., 2020), thus allowing for a mapping to a more theoretically grounded and interlinguistically consistent vocabulary.

Furthermore, we introduce custom tables and columns, not defined in the Paralex standard, but required by the characteristics of our data. In the forms table, an additional column for flexemes is needed, to allow for the expression of both the layers of lexical units described in Section 1. Consequently, an additional table (1d) is also introduced to provide information on flexemes. Inflection classes are assigned to flexemes (rather than lexemes), as lexical items identified according to their form (rather than their meaning) appear to be the appropriate locus to encode a classification based on form. Each flexeme is associated both to its traditional conjugation, expressed with the labels used in the LiLa Knowledge Base of interoperable resources for Latin (Passarotti et al., 2020) – on which see below, Section 3 – and to an index corresponding to its fine-grained inflection micro-class. Micro-classes are automatically inferred from data, using the Qumin toolkit (Beniamine, 2018) to extract alternation patterns between all the possible combinations of principal parts for each flexeme, and grouping together flexemes that share the same set of patterns, as documented in the tables in (1e-f).

### 3 Conversion to RDF and linking to the LiLa Knowledge Base

Having PrinParLat released in the Paralex standard format will make it interoperable with other Paralex lexicons. However, for Latin a wealth of other resources of different kinds is also available, and some of them provide pieces of information that can integrate the ones explicitly recorded in our resource. For instance, in large textual resources like the LASLA corpus (Denooz, 2004) we can find information on the frequencies of wordforms, which is particularly useful as they are generated regardless of their actual attestation in our resource. Lexical resources focusing on other topics can prove useful as well: e.g., a derivation lexicon like Word Formation Latin (Litta & Passarotti, 2020) can give us information on which of the items of our resource are linked by a word formation relation, and how this influences their inflectional behaviour.

To guarantee interoperability with such resources, a richer integration is needed, that can be achieved by means of Semantic Web technologies and standards. Indeed, many of the resources available for Latin have already been made interoperable by connecting them to the LiLa Knowledge Base (cf. Section 2), that follows the RDF data model, where knowledge is

(a) The forms table				
form_id	lexeme	cell	orth_form	flexeme
192	a0105	prs.act.inf	ablauare	a0105
193	a0105	prs.act.inf	ablauere	a0105_2
190	a0105	fut.act.ind.3.sg	ablauabit	a0105
191	a0105	fut.act.ind.3.sg	ablauet	a0105_2

(b) The cells table		
cell_id	unimorph	
prs.act.inf	V;NFIN;ACT;IPFV	

(c) The features-values table			
value_id	value_label	feature	
prs	present	tense-aspect	
act	active	voice	
inf	infinitive	verbform	

(d) The flexemes table			
flexeme_id	inflection_class	lila:Lemma	lila:hasInflectionType
a0105	20	86938	v3r
a0105_2	21	86939	v1r

(e) Mapping inflection classes-patterns			(f) The patterns table			
id	inflection_class	pattern	pattern_id	pattern_alteration	cell_left	cell_right
113	20	1	0	re $\rightleftharpoons$ bit	prs.act.inf	fut.act.ind.3.sg
116	21	0	1	er_ $\rightleftharpoons$ _t	prs.act.inf	fut.act.ind.3.sg

Table 1: The data of PrinParLat

represented in terms of triples that connect a “subject” to an “object” through a “property”, items (“individuals”) are assigned to “classes”, and sub-class and sub-property relations are established between them to describe their characteristics. Following the principles of the Linguistic Linked Open Data paradigm (Cimiano et al., 2020), already existing vocabularies are reused whenever possible – e.g., the OntoLex-Lemon model for lexical resources (McGrae et al., 2017). New classes and properties are introduced whenever necessary. Among the extensions of the LiLa ontology, the crucial one is the class `lila:Lemma`, defined as a subclass of `ontolex:Form`: the core of the Knowledge Base is the Lemma Bank, a large collection of citation forms of Latin words; interoperability is achieved by linking tokens of textual resources and entries of lexical resources to their citation form.

To make our resource interoperable with those already included in the Knowledge Base, we need to also release it in RDF. To do that, Paralex also provides an ontology where tables and columns defined in the standard are mapped to RDF classes and properties, respectively. However, we also need i) to extend this vocabulary to be able to model tables and columns of our resource that are not defined in the standard, and ii) to specify how the conversion should be implemented. Linking to the Knowledge Base can then be performed by connecting each flexeme to its `lila:Lemma` in the Lemma Bank, as shown in the flexemes table in (1d).

## 4 Conclusions and future work

PrinParLat lists the principal parts of Latin verbal (f)lexemes and provides fine-grained information on their inflectional behaviour. Putting all these pieces of information together, it is straightforward to obtain a full lexicon listing all the inflected wordforms of Latin verbs, by performing simple string replacements compatible with the relevant inflection (micro-)class in each of the other cells. The instructions to obtain it can be coded in RDF too, using the vocabulary of the emerging module for the treatment of morphological information in OntoLex lexicons (Chiarcos et al., 2022). Wide-scope interoperability of such a lexicon would be guaranteed with i) other paradigmatic lexicons, thanks to the adoption of the Paralex standard format; ii) other lexical resources that use the OntoLex vocabulary, thanks to the explicit mapping between the Paralex standard and OntoLex provided in the Paralex ontology; iii) resources of

other kind (e.g. corpora), thanks to its release as RDF data linked to the LiLa Knowledge Base.

## References

- Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*, vol. 22. MIT press.
- Beniamine, Sacha. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*: Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7) dissertation.
- Beniamine, Sacha, Olivier Bonami & Benoît Sagot. 2017. Inferring inflection classes with description length. *Journal of Language Modelling* 5(3). 465–525.
- Blevins, James P. 2016. *Word and paradigm morphology*. Oxford University Press.
- Bonami, Olivier & Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182.
- Bonami, Olivier & Berthold Crysmann. 2018. Lexeme and flexeme in a formal theory of grammar. In *The lexeme in descriptive and theoretical morphology*, 175–202. Language Science Press.
- Chiarcos, Christian, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti & Matteo Pellegrini. 2022. Computational Morphology with OntoLex-Morph. In *Proceedings of the 8th workshop on linked data in linguistics within the 13th language resources and evaluation conference*, 78–86.
- Cimiano, Philipp, Christian Chiarcos, John P McCrae & Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.
- Denooz, Joseph. 2004. Opera Latina: une base de données sur internet. *Euphrosyne* 32. 79–88.
- Dressler, Wolfgang U. 2002. Latin inflection classes. In *Theory and description in Latin linguistics*, 91–110. Brill.
- Fradin, Bernard & Françoise Kerleroux. 2003. Troubles with lexemes. In *Topics in Morphology. Selected papers from the Third Mediterranean Morphology Meeting*, 177–196. IULA-Universitat Pompeu Fabra (Barcelona).
- Litta, Eleonora & Marco Passarotti. 2020. (When) inflection needs derivation: a word formation lexicon for Latin. In *Lemmata Linguistica Latina. Words and Sounds*, 224–239. De Gruyter.
- McCarthy, Arya D, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th language resources and evaluation conference*, 3922–3931. European Language Resources Association.
- McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar & Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, 19–21.
- Passarotti, Marco, Marco Budassi, Eleonora Litta & Paolo Ruffolo. 2017. The Lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 workshop on processing historical language*, 24–31.
- Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo & Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici* 58(1). 177–212.
- Stump, Gregory & Raphael A Finkel. 2013. *Morphological typology: From word to paradigm*, vol. 138. Cambridge University Press.
- Thornton, Anna M. 2018. Troubles with flexemes. In *The lexeme in descriptive and theoretical morphology*, 303–321. Language Science Press.

---

# The role of paradigm-external anchoring in simulating the emergence of inflection class systems

*Erich R. Round*

Surrey Morphology Group; Univ. of Queensland

*Sacha Beniamine*

Surrey Morphology Group

*Louise Esher*

CNRS LLACAN

---

## 1 Introduction

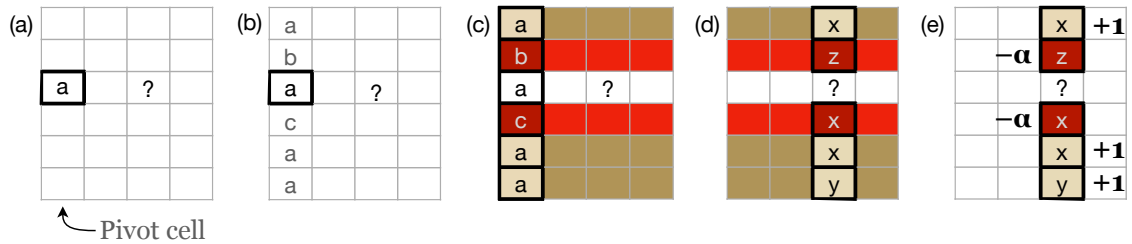
Abstract models of paradigm change can provide insight into how even the simplest processes can lead to unexpected outcomes, thereby revealing new potential explanations for observed linguistic phenomena. Ackerman & Malouf (2015) presented a seminal model in which inflectional systems reduce in disorder through the action of piecemeal analogical changes. More recently, Round et al. (2021a,b) showed that (1) the model cannot evolve stable, persistent inflection classes, rather all inflection classes inevitably collapse together; but (2) if ‘negative evidence’, i.e., evidence from inflectional dissimilarity, is factored into analogical inference, the result can be an attraction-repulsion dynamic which enables inflection classes to coalesce (due to attraction) but remain distinct and thus persist without collapsing together entirely (due to repulsion). The cognitive motivation for attending to negative evidence is presented in Round (forthcoming).

Here we investigate the potential of a third force in inflection class evolution: the anchoring of inflection classes to paradigm-external properties (such as lexical semantics or stem phonology). We compare two conceptions of paradigm-external anchoring, (1) as an analogical force which enters into micro-scale competition with paradigm-internal analogy; and (2) as a soft condition upon paradigm-internal analogy, making some potential paradigm-internal analogies more salient than others, expanding on results by Round et al. (2022).

## 2 A model of analogical change via paradigm cell filling

Inflection classes are sets of lexemes that share inflectional exponents, a type of ‘morphomic’, morphology-internal structure, which mediates the mapping between content and form in inflection (Aronoff, 1994; Round, 2015). The constrained organisation of inflection classes appears to limit the complexity of the inflectional system for language users by offering a systematic, recurrent and predictable means of distributing exponents (Carstairs-McCarthy, 2010; Ackerman & Malouf, 2013; Blevins, 2016; Bonami & Beniamine, 2016). However, a matter of ongoing debate is what kind of historical dynamics could potentially lead to such structure (Carstairs-McCarthy, 2010; Esher, 2015; Maiden, 2018). In this paper, we use computational evolutionary models to shed further light on some of the simplest conditions under which persistent and stable inflection class systems can emerge. These insights will contribute both to theoretical investigations into morphomic structure and to the development of increasingly elaborate models of paradigm evolution in future.

We start from the model of Round et al. (2021a,b), which in turn builds on Ackerman et al. (2009) to model inflectional change via a simple mechanism of paradigm cell filling (PCF). The initial input to the model consists of a lexicon in which paradigms are populated with randomly distributed exponents. The PCF process (Figure 1) is as follows: at each cycle, the model must predict a held out value, termed the *focus* (marked ‘?’ in Fig. 1) at the intersection of a focal cell and focal lexeme. To predict the value of the focus, the model (i) picks a non-focal cell, termed the *pivot* (Fig. 1a); (ii) scans the exponents in pivot cells of other lexemes, termed



evidence lexemes (Fig. 1b), classifying them as possessing a ‘matching’ (green in Fig. 1c) or ‘contrasting’ (red in Fig. 1c) pivot compared to the focal lexeme; (iii) inspects the exponents of focal cells in the evidence lexemes (Fig. 1d) and scores exponents of matching lexemes positively (as +1) and of contrasting lexemes negatively (as  $-\alpha$ , where  $\alpha$  is a non-negative value set by the experimenter); (iv) sums the scores for each exponent type in the focal cell and selects the highest-scoring for the focal exponent. Round et al. (2021a,b) show that when only positive evidence is taken into account, i.e., when  $\alpha = 0$ , inflection classes will invariably collapse together as the PCF process continues to repeat, whereas when negative evidence is attended to, i.e.,  $\alpha > 0$ , it is possible for distinct inflection classes to emerge and persist stably.

### 3 Adding sensitivity to paradigm-external properties

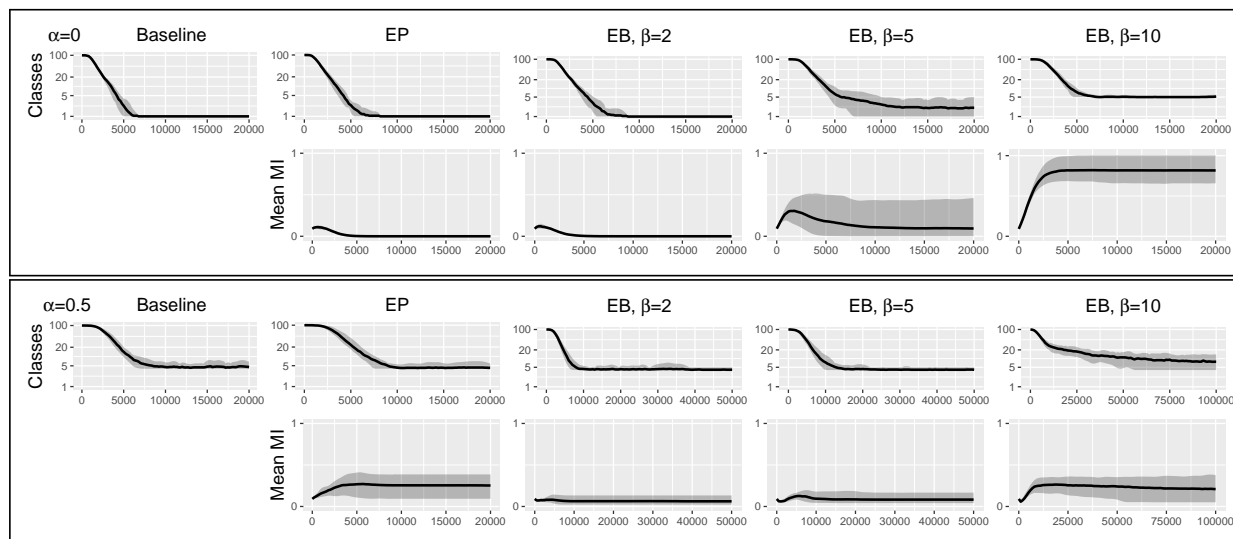
Here we ask whether persistent inflection classes can also emerge via a second mechanism. In the model of Round et al. (2021a,b), the PCF process is affected solely by the inflectional exponents. However analogical change and inflectional predictability are known to be influenced by other lexical properties (Guzmán Naranjo, 2018). We therefore enrich the model, to investigate the contribution of such properties to inflection class emergence. For each lexeme, we add one more discretely-valued property—effectively one more column in Figure 1—which can be interpreted as any non-paradigmatic lexical property, for instance semantic (animate v. inanimate nouns), syntactic (intransitive v. transitive v. ditransitive verbs) or phonological (types of stems).

We test two ways in which paradigm-external information could contribute to PCF. The PCF mechanism will never change the paradigm-external properties of lexemes. In the first, ‘External Pivot’ (EP) model, paradigm-external properties can be selected as the pivot, in which case their similarities and differences determine whether evidence lexemes are classified as ‘matching’ or ‘contrasting’. In the second, ‘External bias’ (EB) model, the paradigm-external properties are never pivots, thus the ‘matching/contrasting’ classification is determined solely by inflectional exponents, however, evidence lexemes whose paradigm-external property matches that of the focal lexeme receive an enhanced salience, which is implemented within the scoring procedure (Figure 1e) by multiplying the scores of these evidence lexemes by a weighting multiplier,  $\beta$ .

### 4 Results and discussion

We present simulations without negative evidence ( $\alpha = 0$ ) and with it ( $\alpha = 0.5$ ), using 3 models: the Baseline model (Round et al., 2021a,b), the External Pivot model and the External Bias model (with  $\beta \in \{2, 5, 10\}$ ). Inflectional systems contained 100 lexemes with 8 cells. For comparability with earlier models (§1), each cell and also the paradigm-external property had 5 possible exponents/values, initialised randomly. The PCF process was then iterated and the evolution of

Figure 2: Simulation Results. Upper panel without negative evidence ( $\alpha = 0$ ); Lower panel with it ( $\alpha = 0.5$ ). In each panel, Upper row: number of classes (plotted on a nonlinear scale); Lower row: mean MI between paradigm cells and the extra-paradigmatic property. Horizontal axis: number of PCF cycles. Black lines show means, grey ribbons 80% of the variation, for 20 repetitions. Models are: Baseline, External Pivot (EP) and External Bias (EB; with  $\beta \in \{2, 5, 10\}$ ).



the system measured in terms of the number of distinct inflection classes and the mean of the mutual information (MI) between each inflectional cell and the external property. Results appear in Figure 2. We comment on classes, then mean MI.

Without negative evidence, the EP model shows negligible difference from the baseline. All inflection classes collapse together. This was true even when we biased the model to select the paradigm-external properties as the pivot 10 $\times$  as frequently as paradigm cells. Thus we conclude that taking paradigm-external properties as pivots in analogical change does not promote the emergence of persistent inflection classes. In contrast, the EB model does lead to the emergence of 4 or 5 persistent inflection classes when  $\beta = 10$ , and to classes that are long-lived—though fewer in number, and not indefinitely persistent—when  $\beta = 5$ . Classes collapse when  $\beta = 2$ . Thus, inflection classes can emerge and persist when the weighting of evidence lexemes is strongly biased (but not when weakly biased) on the basis of paradigm-external similarity.

With negative evidence, 4 or 5 inflection classes reliably emerge, even in the baseline model. The EP model shows negligible difference from the baseline. This indicates again that when external properties function as analogical pivots, they have little impact. Results from the EB model are similar to those from the EP model when  $\beta = 2$  and  $\beta = 5$ . However, when  $\beta = 10$ , a more interesting dynamic plays out. Inspection of the classes reveals that while 4 or 5 classes slowly emerge, they pass through early stages containing multiple micro-classes (small variations on the theme), with the micro-classes aligning closely with the external properties. However, the micro-classes subsequently collapse, eventually leaving only the 4 or 5 macro-classes.

When inflection classes emerge, mean MI in Figure 2 indicates the degree to which the class memberships of lexemes align with their paradigm-external properties. (When the classes collapse, MI trivially goes to zero.) Here we see that although stable classes arise without negative evidence ( $\alpha = 0$ ) in the EB model with strong bias ( $\beta = 10$ ), those classes are distributed in the lexicon in very close alignment with lexemes' external properties, and thus exhibit little 'morphological autonomy' (Aronoff, 1994). Only in models with negative evidence do the emergent classes diverge significantly from the paradigm-external properties.



## 5 Conclusions

Round et al. (2021a,b) showed that with negative evidence, piecemeal analogical changes can lead to emergent and persistent inflectional classes. Here we confirm that paradigm-external anchoring can also do so, but in the absence of negative evidence, the emergent classes exhibit little morphological autonomy. This suggests that negative evidence may be indispensable for the emergence of truly autonomous inflection classes. We find that when external anchoring is combined with negative evidence, a mix of autonomous and non-autonomous patterning emerges, much as in real inflection class systems (e.g. Bybee & Moder, 1983). This dynamic interaction is a new and fascinating finding that warrants further investigation.

## References

- Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar*, 54–82. Oxford: OUP.
- Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429–464.
- Ackerman, Farrell & Robert Malouf. 2015. The No Blur Principle Effects as an Emergent Property of Language Systems. In *Berkeley Linguistics Society* 41, doi:10.20354/B4414110014.
- Aronoff, Mark. 1994. *Morphology by itself*. Cambridge: MIT Press.
- Blevins, James P. 2016. *Word and Paradigm Morphology*. Oxford: OUP.
- Bonami, Olivier & S. Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182. doi:10.3366/word.2016.0092.
- Bybee, Joan L & Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language* 59(2). 251–270.
- Carstairs-McCarthy, Andrew. 2010. *The Evolution of Morphology*. Oxford: OUP.
- Esher, Louise. 2015. Approaches to the function and emergence of metamorphomic structure. *Décembrettes* 9, Toulouse.
- Guzmán Naranjo, Matías. 2018. *Analogical classification in formal grammar*. Berlin: Language Science Press.
- Maiden, Martin. 2018. *The Romance verb: Morphomic structure and diachrony*. Oxford: OUP.
- Round, Erich, Sacha Beniamine, Louise Esher, Matt Spike et al. 2022. Cognition and the stability of evolving complex morphology: an agent-based model. In *Proceedings of the Joint Conference on Language Evolution (JCoLE)*, 635–642. JCoLE.
- Round, Erich R. 2015. Rhizomorphomes, meromorphomes, and metamorphomes. In Greville Corbett, Dunstan Brown & Matthew Baerman (eds.), *Understanding and Measuring Morphological Complexity*, 29–52. Oxford: OUP.
- Round, Erich R. forthcoming. Models, morphomes, emergence and the mind. In Xavier Bach, Louise Esher & Sascha Gaglia (eds.), *Comparative and dialectal approaches to analogy: Inflection in romance and beyond*, Oxford: OUP.
- Round, Erich R., Sacha Beniamine & Louise Esher. 2021a. The role of attraction-repulsion dynamics in simulating the emergence of inflectional class systems. *International Symposium of Morphology 2021*, <https://doi.org/10.48550/arXiv.2111.08465>.
- Round, Erich R., Sacha Beniamine & Louise Esher. 2021b. Spontaneous emergence of inflectional class systems via attraction-repulsion dynamics. *American International Morphology Meeting 2021*, <https://www.ling.ohio-state.edu/AIMM5/schedule.html>.

---

# Innovative uses of French neological *-ance* nominalizations

*Philippe Gréa*  
Université Paris Nanterre &  
Modyco, UMR 7114

*Marie L. Knittel*  
Université de Lorraine & ATILF,  
UMR 7118

*Rafael Marín*  
CNRS, STL, UMR 8163 &  
Université de Lille

*Florence Villoing*  
Université Paris Nanterre &  
Modyco, UMR 7114

---

## 1 Data

French *-ance* nominals (N-*ance*) are mostly built on verbs (*surveiller* ‘to monitor’ > *surveillance* ‘monitoring’) and adjectives (*élégant* ‘elegant’ > *élégance* ‘elegance’) (Dal & Namer 2010, Knittel 2016).

However, in the standard lexicon, Object Experiencer Psychological Verbs (OEPV) are not used as basis for *-ance* nominalizations (Knittel 2016, Knittel & Marín 2022, from the nouns registered in the lexical base Lexique3 (New et al. 2001). The only example found in this base, *attirance* ‘attraction’ from *attirer* ‘to attract’, has been coined by Baudelaire (Rey & al. 1998). This peculiarity sets *-ance* deverbal nominalization apart from *-ion* (1a) and *-ment* (1b) suffixation and V to N conversion (1c) nominalization patterns, that are commonly used with this verb class (Barque, Fábregas & Marín 2012).

- (1) a. *fasciner* > *fascination* ; *obséder* > *obsession* ; *préoccuper* > *préoccupation*  
to fascinate / fascination; to obsess / obsession; to preoccupy / preoccupation  
b. *apaiser* > *apaisement* ; *décourager* > *découragement* ; *épanouir* > *épanouissement*  
to appease / appeasement; to discourage / discouragement; to fulfill / fulfilment  
c. *craindre* > *crainte* ; *désirer* > *désir* ; *regretter* > *regret*  
to fear / fear; to desire / desire; to regret / regret

Yet, at first sight, *-ance* neological nominalizations are frequently related to OEPV (2a). Furthermore, they regularly belong to morphological families also comprising an *-ant* adjective built on the verb, thus resulting in triplets (2b).

- (2) a. *charmer* / *charmance* ; *écoeurer* / *écoeurance* ; *désoler* / *désolance*  
to charm / charm-ance to disgust / disgust-ance to afflict/afflict-ance  
b. *charmer* / *charmant* / *charmance* ; *écoeurer* / *écoeurant* / *écoeurance* ;  
to charm / charming / charm-ance to disgust / disgusting / disgust-ance  
*désoler* / *désolant* / *désolance*  
to afflict / afflicting / afflict-ance

## 2 The issue

The question is why *-ance* can form neological nominalizations from OEPV, contrary to what is observed in the general lexicon.

We show that depending on their interpretation, and the inheritance of the arguments of the base verb, *-ance* nominals are either built on the verb or on the *-ant* adjective.

## 3 The database

The database on which our study is based comprises 350 neological nominals built with the nominalizing suffix *-ance*, extracted from the frCow, which is currently the largest and more recent web corpus available for French (it comprises 9 billion words extracted from the Internet (Schäfer & Bildhauer 2012; Schäfer 2015), lemmatized, annotated according to their syntactic categories and informed by frequency (Missud, Amsili, Villoing 2020).

We based the selection of neological N-ance on their frequency of occurrence in corpora. Only those having a 1 to 20 frequency have been selected. We identified this frequency as optimal to detect neologisms; on the one hand, we have enough varied contexts to grasp the meaning and analyze the environment of the neological form; on the other hand, we observed that nouns with higher frequencies are often not neological, and belong to specialized vocabulary. The data have first been automatically processed, so as to constitute plausible V-N pairs, then sorted manually. Contexts of discursive use extracted from Google or Twitter, were then added for each N-ance, in order to grasp their meanings. At the end of this process, about 350 neological -ance nominals were gathered, with contexts of use, and paired with the morphologically related adjectives (273 adjectives) and verbs (322 verbs). Finally, the base verbs were annotated according to their lexical aspect, and the -ance nominals for their argument structure. Among these 322 verbs, we detected 51 OEPV, which represents 15,83% of the verbs paired with neological N-ance. We can thus count 51 trio of N-ance / OEPV / ADJ-ant on OEPV.

## 4 Results

The argument structure of OEPV is presented in (3) (Arad, 1998; Pesetsky, 1995; Pylkkänen, 2000). When such verbs are nominalized by *-ment* / *-ion* suffixation, the corresponding noun inherits the Experiencer argument of their base verbs, which is introduced by *de* 'of' (Grimshaw, 1990). The Stimulus is optional and surfaces as a PP introduced by *pour* 'about' / 'with', and less frequently by *par* 'by' (4).

- (3) a. Subject<sub>STIMULUS</sub> V Object<sub>EXPERIENCER</sub>  
 b. {*Pierre / la musique*}<sub>Stim</sub> {*fascine / émerveille / apaise / dérange*} Marie<sub>Exp</sub>.  
 'Pierre / music} {fascinates / delights / appeases / bothers} Marie.'
- (4) a. *la* {*fascination / émerveillement*} *de* Marie<sub>Exp</sub> *pour* {*Pierre / la musique*}<sub>Stim</sub>.  
 'the fascination / delighting} of Marie with {Pierre / music}.'  
 b. *l'*{*apaisement / dérangement*} *de* Marie<sub>Exp</sub> *par* {*Pierre / la musique*}<sub>Stim</sub>.  
 'the {appeasement / bothering} of Marie by {Pierre / music}.'

The data we gathered show that, by contrast, a large part of neological N-ance built on OEPV inherit the Stimulus argument (5), a pattern that is not available for lexicalized N-ance.

- (5) a. *l'apaisance du reggae*<sub>STIM</sub> ≈ '*Le reggae*<sub>STIM</sub> *est apaisant.*'  
 lit. 'the appease-ance of reggae' ≈ 'Reggae is appeasing.'  
 b. *la déconcertance du mec*<sub>STIM</sub> (*à ce sujet*) ≈ '*Ce mec*<sub>STIM</sub> *est déconcertant.*'  
 lit. 'the puzzle-ance of the guy (on that matter)' ≈ 'This guy is puzzling.'

Yet, they can also inherit the Experiencer, also realized by a *de* PP (6).

- (6) *la fascinace des Targaryen*<sub>Exp</sub> *pour les dragons*<sub>STIM</sub>  
 'the fascin-ance of the Targaryans with dragons'

These data raise the question of the origin of this uncommon inheritance pattern.

Two competing hypotheses can be suggested.

- i. The -ance suffix exhibits a particular behavior in neologisms, in that it enables the inheritance of the Stimulus argument of the verb. However, this hypothesis is highly improbable, since no model predicts that -ance can behave differently from the general lexicon.
- ii. The base of -ance nominal is not the verb, but the corresponding adjective -ant adjective, also derived from the corresponding OEPV.

Two arguments favor the second hypothesis.

First, -ant adjectives can be used as bases of -ance nominals in neologisms, cf. *méchance*<sub>N</sub> 'wickedness' from *méchant*<sub>Adj</sub> 'wicked', as well as in the general lexicon, cf. *élégance* 'elegance' from *élégant* 'elegant'. In the absence of a verb, these -ance nominals can only be built on adjectives.

Second, when a neological N-*ance* inherits the Stimulus argument of the verb, it regularly behaves as a property-denoting nominal, and typically refers to an inherent property (or quality) of an individual. This is why, unlike event and state nouns, they are compatible with the so-called genitive of quality, intensity markers, and the expression of paragon (Flaux & Van de Velde, 2000). This is indeed the case for the neological nouns *contraingance* lit. 'coerce-ness', *époustouflance* lit. 'stuning-ness', and *gênance* lit. 'bother-ness'.

- |   |                       |
|---|-----------------------|
| (7) a. <i>Cet accord est d'une contraingance ridicule</i> | [genitive of quality] |
| lit. 'This agreement is of ridiculous coerce-ance'        |                       |
| b. <i>l'époustouflance absolue du design</i>              | [intensity]           |
| lit. 'the absolute astonish-ance of the design'           |                       |
| c. <i>le prime de la gênance</i>                          | [parangon]            |
| lit. 'the height of bother-ance'                          |                       |

Crucially, according to Flaux & Van de Velde (2000), derived property nominals are mostly adjectival based.

Thus, our analysis suggests that all neological *-ance* nominals that inherit the Stimulus argument are built on adjectival bases, that are in turn built on verbs. The Stimulus argument is in fact inherited by the adjective, and transmitted to the corresponding *-ance* nominal when the adjective is nominalized.

## 5 Conclusion

To conclude, our analysis of neological N-*ance* has shown, on the one hand, a new tendency of deverbal suffixation in *-ance* to take OEPV as bases similarly to *-ment* and *-ion* nominalizations, whereas this possibility has not yet been exploited by *-ance* nominalizations in the standard lexicon. On the other hand, we have shown that there are two construction patterns available for N-*ance* neologisms that have an OEPV in their morphological family:

- the first has a verbal base. In this case, only the Experiencer argument is inherited (6). It also reveals the originality of *-ance* neological suffixation, that can maintain the Experiencer argument, similarly to *-ment* and *-ion* nominalizations, a pattern that is however less frequent.
- the second has an adjective as a base, the adjective itself being deverbal (5). The *-ance* nominal does not inherit the stimulus argument from the verb, but from the adjective.

This double construction is enabled by the fact that *-ance* suffixation can select either verbal or adjectival bases.

We therefore observe, in the line of Dal & Namer (2010), the facilitating character of a morphological family containing a verb and a related adjective in *-ant* for the emergence of a N-*ance*. However, contrary to what they stated for lexicalized N-*ance*, it is not secondary to decide whether the noun is built on the verb or on the adjective in *-ant*, since the category of the base determines which argument is inherited by the N-*ance*. If the derivational family is indeed a facilitator in the emergence of an N-*ance* related to an OEPV, this is related to the existence of binary relations between the members of this family: between verbs (OEPV) and *-ant* adjectives, and between *-ant* adjectives and *-ance* nouns.

There is one question, however, that remains to be properly addressed: why we found *-ance* neological nominalizations related to OEPVs, contrary to what is observed in the general lexicon? Part of the answer would be that psych nouns in standard French, mostly derived from verbal bases (typically with *-ment* and *-ion* suffixes), systematically denote states (Barque et al., 2012). By contrast, neological *-ance* suffixation has the capacity to generate psych nouns (from *-ant* adjectives) denoting qualities. This is, we argue, the gap that many speakers try to fill, in a similar way as Charles Baudelaire did in his time, quite successfully, with his innovative *attirance du gouffre* 'attraction of the abyss'.

## References

- Arad M. 1998. Psych-notes. *UCL Working Papers in Linguistics* 10.
- Barque L., Fábregas A. & Marín, R. 2012. Les noms d'états psychologiques et leur "objet" : étude d'une alternance sémantique. *Lexique* 20 (2012). 21-41.
- Dal G. 2004. *Vers une morphologie de l'évidence : d'une morphologie de l'input à une morphologie de l'output*. Doctoral dissertation, Université Lille 3.
- Dal G., Namer F. 2010. Les noms en -ance/-ence du français : quels patrons constructionnels ? In F. Neveu et al. (eds.), *Actes en ligne du 2e Congrès Mondial de Linguistique Française*, 893-907. La Nouvelle Orléans. 12-15 juillet 2010.
- Flaux N., Van de Velde V. 2000. *Les noms en français. Esquisse de classement*. Ophrys, Paris.
- Grimshaw J. 1990. *Argument structure*. Cambridge MA: MIT Press.
- Knittel M.L. 2016. Les noms en -ance : un panorama. In F. Neveu, G. Bergounioux, M-H Côté, J.-M. Fournier, L. Hriba et S. Prevost (eds.), *Actes du CMLF 2016*. ILF.
- Knittel M.L. & Marín R. 2022. L'héritage transcatégoriel des propriétés sémantiques : le cas des noms en -ance et des verbes et adjectifs apparentés. Actes du CMLF 2022, F. Neveu, S. Prévost, A. Steuckardt, G. Bergounioux and B. Hamma (eds.), ILF.
- Lignon S., F. Namer et F. Villoing. 2014. De l'agglutination à la triangulation ou comment expliquer certaines séries morphologiques. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer & S. Prévost (eds.), *Actes du 4ème Congrès Mondial de Linguistique Française*, 1813-1835. ILF, Berlin.
- Missud A., Amsili, P. & Villoing, F. 2020. VerNom : une base de paires morphologiques acquise sur très gros corpus. *Actes de la 27e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2020)*, 305-313.
- Missud, A., Villoing, F. 2022. Nominalisations sans base verbale suffixes en -ion, -age et -ment du français : conditions morphologiques. Actes du 8ème Congrès Mondial de Linguistique Française (CMLF 2022), Orléans 4-8 juillet 2022, Paris EDP Sciences.
- Namer F. 2009. *Morphologie, lexique et traitement automatique des langues*. Hermès-Lavoisier.
- New B., Pallier C., Ferrand L., Matos R. 2001. Une base de données lexicales du français contemporain sur internet : LEXIQUE, *L'Année Psychologique*, 101, 447-462. <http://www.lexique.org>.
- Pesetsky D. 1995. *Zero Syntax: Experiences and Cascades*. Cambridge MA: The MIT Press.
- Pylkkänen L. 2000. On stativity and causation. In C. Tenny & J. (eds.), *Events as Grammatical Objects*.
- Rey A. & al. 1998. *Dictionnaire Historique de la langue française*. Paris: Dictionnaires Le Robert
- Schäfer R. 2015. Processing and querying large web corpora with the COW14 architecture. In P. Baaski, H. Biber, E. Breitnender, M. Kupietz, H. Langen & A. Witt (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Lancaster: UCREL IDS.
- Schäfer R. & Bildhauer, F. 2012. Building large corpora from the web using a new efficient tool chain. In n. C. C. Chair, K. Choukri, T. Declerck, M. U. Doăyan, B. Maegaard, J. Mariani, A. Moreno, J. Odiijk & S. Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 486-493. Istanbul, Turkey: European Language Resources Association (ELRA).

---

# On the analysis of the neological resultative construction suffixed with *-iser* and *-化* [-huà] in contemporary media

Jiahui ZHU

Université Sorbonne Nouvelle & Lattice (CNRS/ENS-PSL/USN)

---

## 1 Introduction

Both the French suffix *-iser* and the Chinese suffix *-化* [-huà] have the potential to construct the meaning of “make <patient> become a state” (Fradin, 2003: 296; Zhang & Song, 2007 : 105). In other words, it is reasonable to conjecture that these two verbal suffixes are able to lead and form resultative constructions within the framework of the *Construction Grammar* (e.g. Goldberg, 1995; Goldberg & Jackendoff, 2004). In the post-2000s, with the popularity of the Internet, a large number of neologisms conforming to the structures [[X]<sub>Adj./N.</sub> -iser]<sub>v.</sub> and [[X]<sub>Adj./N.</sub> -化 [-huà]]<sub>v.</sub> such as *googliser*, *neymariser*, 精准化 ([jīng zhǔn huà]), 微信化 ([wēi xìn huà]) and *etc.*, enter the community. Thus, *which resultative sub-constructions are interpreted by these verbal suffixed neologisms in French and in Chinese? Can -iser and -化 [-huà] interpret the same resultative sub-constructions?* In order to answer these questions, guided by the *Cognitive Construction Grammar* (e.g. Goldberg, 1995; Bouveret & Legallois, 2012) and *Construction Morphology* (e.g. Booij, 2010 ; Booij & Audring, 2017), we qualitatively study verbal neologisms suffixed with *-iser* and *-化* [huà] appearing and diffusing after the year 2000 in media discourse and identify the neological resultative sub-constructions explained by these two suffixes. The result shows that *-iser* and *-化* [-huà] are able to interpret the same resultative sub-constructions.

## 2 Data

The neologisms allow us to understand the derivational process in its most regular form, without the semantic opacity that occurs with lexicalization (Huyghe & Lombard, 2022 : 25). In order to show the relationship between French resultative constructions suffixed with *-iser* and Chinese ones suffixed with *-化* [huà], based on written media texts from 2000 to 2022, we construct a corpus of the verbal neologisms derived by these two suffixes. The Chinese metadata are from the traditional media *People's Daily*<sup>1</sup>, *Nanfang Daily*<sup>2</sup>, and the modern media *Sina weibo*<sup>3</sup>; while the French metadata comes from both traditional and new media: all French newspapers on the *Europresse*<sup>4</sup> platform and tweets from *Twitter*<sup>5</sup> (e.g. Daoust, 2017).

More precisely, the constitution of the corpus can be described as follows. Firstly, thanks to the *Jieba* (Sun, 2012) and *Spacy* (Honnibal & Moutani, 2017), two Python machine learning libraries, we automatically select all French terms ending in the morpheme *-iser* and its inflections, and Chinese terms ending in the character *-化* [huà]. In our metadata, we extract 1427 Chinese terms and 2627 French terms. Secondly, we semi-automatically delete terms that are irrelevant to our research. On the one hand, we semi-automatically eliminate faulty forms, non-suffixed forms, suffixed non-verbal terms, and hapaxes separately. On the other

---

<sup>1</sup><http://www.people.com.cn>

<sup>2</sup><https://epaper.southcn.com>

<sup>3</sup><https://m.weibo.cn>

<sup>4</sup><https://nouveau.europresse.com>

<sup>5</sup><https://twitter.com>

hand, we remove terms that appear and diffuse before the year 2000. For French, by referring to the *Google Ngram* platform (Lin *et al.*, 2012) ([1500-2019]), the *Europresse* platform ([1840 to present]), and new media *Twitter* ([2006 to present]), we date the first appearance of the terms. For Chinese, using the press platforms *Modern Newspaper in China*<sup>6</sup> ([1840-1949]); *Chinese Digital Library*<sup>7</sup> ([1946 to present]); *China Core Newspapers Full-text Database*<sup>8</sup> ([2000 to present]) and the new media: *Sina Weibo* ([2009 to present]), we determine the date of the first appearance of the terms. Then we delete the terms that appear and spread before the year 2000. The terms that were the hapaxes before the year 2000 but are largely diffused after 2000, are also included in this study. Finally, we select a total of 1,200 French and 700 Chinese verbal neologisms. Thanks to the concordancer TXM (Heiden *et al.*, 2010), we extract 24000 French concordances and 9278 Chinese concordances associated with each construction. By virtue of the systematical studies of the corpus, we identify the resultative sub-constructions interpreted by these two suffixes in Chinese and French.

### 3 Analysis

In order to analyse and identify the resultative sub-constructions, we study separately the semantic and syntactic roles of the roots in the resultative construction. The sub-constructions are classified in terms of the different syntactic-semantic roles of the roots.

According to the semantic aspect of the constructional approach, the interpretation of the resultative construction has two sub-events: *the verbal sub-event* and *the constructional sub-event* (Goldberg & Jackendoff, 2004 : 538). The *verbal sub-event* can play the role of indicating that the state, that the patient has acquired, is a result of a change caused by an action, rather than their own original state. The *constructional sub-event* plays the role of indicating which state change the agent has made to the patient. In the absence of an agent, the *constructional sub-event* simply denotes which state change the patient has undergone. For the neologisms conforming to the structure  $[[X]_{Adj./N.} -iser]_V$ . and  $[[X]_{Adj./N.} -{化} [-huà]]_V$ , the presence of *-iser* & *-化* [-huà] suggests that such states are not original properties of the patient. Compared with the verbal sub-events expressing manners, the verbal sub-events completed by the suffixes in question cannot clarify the manners in which the patient obtains the new states. The new state obtained in the constructional sub-event is interpreted by the root X. When the agent occupies a place in the constructional sub-event, the phrasal resultative construction directed by the suffixed constructions is a *causative-explicit resultative construction*. On the other hand, when the agent is absent in the constructional sub-event, the phrasal resultative construction directed by the suffixed constructions is a *causative-implicit resultative construction*.

In the morphological aspect, affixes cannot possess meaning at the level of semantics and syntax independently of the derivatives. Thus, in the study of the syntactic role of the roots in the construction, we are not interested in the syntactic relationship between these roots and *-iser* and *-化* [huà]. Rather, from the syntactic point of view, we investigate the position and syntactic role of these roots in the predication of the resultative suffixed construction. Combining this investigation with the classification between *causative-explicit resultative construction* and *causative-implicit resultative construction* (*i.e.* the two sub-categories obtained from the semantic aspect), according to the systematic studies of Chinese and French suffixed neologisms in the constructed corpus (see section 2), we have identified the following sub-categories:

i. When the root plays the role of the predicative in the predication, these suffixes in question form the resultative constructions of the property.

<sup>6</sup><http://tk.cepiec.com.cn>

<sup>7</sup><http://www.apabi.com/jigou?pid=about&cult=CN>

<sup>8</sup><https://kns.cnki.net/kns/brief/result.aspx?dbprefix=CCND>

- (1) Causative-explicit property resultative:<sup>9</sup>
- a. *Il se porte alors avec veste et chemise pour **chiciser** la silhouette.* (Challenges, 23/06/2016)
- b. “税企通” 使 纳税服务 [精准-化]. (Nanfang’s Daily 09/07/2014)  
 “shuìqītōng” shǐ nàshuìfúwù [jīngzhǔn-huà<sub>suff.</sub>]  
 “ShuiQiTong”<sub>n.</sub> POM Tax-Services<sub>n.</sub> [Precise<sub>adj.</sub>-huà<sub>suff.</sub>] v.  
 “ShuiQiTong” makes tax services precise.
- (2) Causative-implicit property resultative:
- a. *Un 49.3 qui **macronise**.* (La Montagne, 24/02/2015)
- b. 人居 环境 [低碳-化]. (People’s Daily, 13/03/2013)  
 rénjū huánjìng [dītàn-huà<sub>suff.</sub>]  
 habitat-human<sub>adj.</sub> Environment<sub>n.</sub> [Low-carbon<sub>n.</sub>-huà<sub>suff.</sub>] v.  
 Habitat becomes low-carbon.
- ii. When the root is not the predicative in the predication, its interpretation is supported by another additional predicate, thus acting as an *oblique complement* (OC) of the predication. Based on the semantic aspect of the “OC”, we have divided this one category into two sub-categories: *recipient resultative* and *means resultative*.
- (3) Causative-explicit recipient resultative :
- a. *Cela est très lié à notre capacité de pouvoir **APIser** notre système pour donner la possibilité de s’y connecter rapidement.* (IT for Business, 13/12/2021)
- b. (他) 使 手术 [微创-化]. (Nanfang’s Daily 09/07/2014)  
 (tā) shǐ shǒushù [wēichuàng-huà<sub>suff.</sub>]  
 (It)<sub>pron.</sub> POM operation<sub>n.</sub> [small<sub>adj.</sub>-incision<sub>n.</sub>-huà<sub>suff.</sub>] v.  
 (It) makes the operation have small incisions.
- (4) Causative-implicit recipient resultative :
- a. *Nous faisons face aux mêmes problématiques que les banques ou les assureurs, avec des systèmes qu’il faut progressivement **APIser**.* (L’Usine Nouvelle, 01/12/2022)
- b. 东方甄选 开始 [去-辉-化]. (Weibo, 08/02/2023)  
 dōngfāngzhēnxuǎn kāishǐ [qù-huī-huà]  
 DongFangZhenXuan<sub>company name</sub> Begin<sub>v.</sub> [pref.<sub>neg.</sub>-Hui<sub>name of person</sub>-huà<sub>suff.</sub>] v.  
 DongFangZhenXuan starts to become without Hui.
- (5) Causative-explicit means resultative :
- a. *Les pays des Balkans ne sauraient **euroïser** leurs économies pour contourner le traité de Maastricht en vue d’une entrée dans l’Union européenne.* (Les Echos, 30/11/2004)
- b. 苏宁 开始 将 销售 业务 [微信-化]. (Sina Weibo, 16/02/2014)  
 sūníng kāishǐ jiāng xiāoshòu yèwù [wēixìn-huà]  
 SuNing<sub>n.</sub> Begin<sub>v.</sub> POM Sell<sub>v.</sub> Business<sub>n.</sub> [Wechat<sub>n.</sub>-huà<sub>suff.</sub>] v.  
 Suning starts using WeChat for sales operations
- (6) Causative-implicit means resultative :
- a. *Je suis transparent, avec une identité facile à **Googliser**.* (Twitter, 19/07/2021)
- b. 汇款 转账 [二维码-化]. (Nanfang’s Daily, 30/01/2013)  
 huìkuǎn zhuǎnzhàng [èrwéimǎ-huà]  
 Remittance<sub>n.</sub> Transfer-of-account<sub>n.</sub> [QR-code<sub>n.</sub>-huà<sub>suff.</sub>] v.  
 Remittances and transfers are available via QR codes.

<sup>9</sup>In these examples, a is a French example and b is a Chinese example, POM = pre-verbal object marker.



Based on the six resultative sub-constructions identified above in French and Chinese, we propose a common schema shown in (7) for the resultative construction suffixed with *-iser* and *-化* [-huà]. The morpho-syntactic and semantic structures are represented from left to right.

(7) <[[X<sub>Adj./N</sub>]<sub>i</sub> -iser/-化 [huà]]<sub>vj</sub>> ↔ <[Cause [patient] to become state relating to SEM<sub>i</sub>]<sub>vj</sub>>.

## 4 Discussion

Based on the constructional approach, this study analysed the syntactic-semantic roles of the roots of neologisms in contemporary media. The result allows us to conclude that there are six resultative sub-constructions suffixed with *-iser* and *-化* [-huà] (see section 3): *causative-explicit property resultative*, *causative-implicit property resultative*, *causative-explicit recipient resultative*, *causative-implicit recipient resultative*, *causative-explicit means resultative* and *causative-implicit means resultative*. The identification of these sub-constructions and the schema shown in (7) allows further comparison of the characteristics of the resultative construction suffixed with *-iser* and *-化* [-huà]. Moreover, this result demonstrates that the notion of construction is valuable and feasible for cross-linguistic comparison and analysis in the morphological aspect.

## References

- Booij, G. 2010. *Construction morphology*. New York: Oxford University Press.
- Booij, G. & J. Audring. 2017. Construction morphology and the parallel architecture of grammar. *Cognitive science* 41. 277–302.
- Bouveret, M. & D. Legallois. 2012. Cognitive linguistics and the notion of construction in french studies. In M. Bouveret & D. Legallois (eds.), *Constructions in french*, 1–19. Amsterdam / Philadelphia: John Benjamins.
- Daoust, J.-F. 2017. Démocratisation de l'information : effets différenciés des médias traditionnels et des nouveaux médias. *Politique et Sociétés* 36(1). 25–46.
- Fradin, B. 2003. *Nouvelle approches en morphologie*. Paris: Presses Universitaires de France.
- Goldberg, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Goldberg, A.E. & R. Jackendoff. 2004. The english resultative as a family of constructions. *Language* 80. 532–568.
- Heiden, S., J. P. Magué & B. Pincemin. 2010. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Sergio Bolasco, Isabella Chiari & Luca Giuliano (eds.), *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, vol. 2 3, 1021–1032. Rome, Italy: Edizioni Universitarie di Lettere Economia Diritto.
- Honnibal, M. & I. Montani. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. To appear.
- Huyghe, R. & A. Lombard. 2022. Les néologismes en *-age* en français contemporain: héritage verbal et polysémie. *Journal of French Language Studies* 32(1). 25–47.
- Lin, Y.-R, J.-B Michel, E. Lieberman Aiden, J. Orwant, W. Brockman & S. Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the 50th annual meeting of the association for computational linguistics, volume 2: Demo papers (acl '12)*, .
- Sun, J.-Y. 2012. Jieba chinese word segmentation tool, *Accessed : Jun 25 : 2018*.
- Zhang, Y. F. & J. Song. 2007. A research of the causative voice sentence with the chinese verbal ending hua. *Fudan Journal (Social Sciences)* 4. 105–110.

